

PHYLOGENETIC SUPERTREES

Computational Biology

VOLUME 4

Editor-in-Chief

Andreas Dress, *University of Bielefeld, Germany*

Editorial Board

Gene Myers, *Celera Genomics, Maryland, U.S.A.*

Robert Giegerich, *University of Bielefeld, Germany*

Walter Fitch, *University of California, Irvine, CA, U.S.A.*

Pavel A. Pevzner, *University of California, Irvine, CA, U.S.A.*

Advisory Board

Gordon Gripper, *University of Michigan*; Joe Felsenstein, *University of Washington*;

Dan Gusfield, *University of California, Davis*; Sorin Istrail, *Sandia National Laboratories*; Samuel Karlin, *Stanford University*; Thomas Lengauer, *GMD-Sankt Augustin, Germany*; Marcella McClure, *Montana State University*; Martin Nowak,

Princeton University; David Sankoff, *University of Montreal*; Ron Shamir,

Tel Aviv University; Mike Steel, *University of Canterbury, New Zealand*;

Gary Stormo, *Washington University Medical School*; Simon Tavaré, *University of Southern California*; Martin Vingron, *DKFZ, Heidelberg*; Tandy Warnow,
University of Texas, Austin

PHYLOGENETIC SUPERTREES

Combining information
to reveal the Tree of Life

Edited by

OLAF R.P. BININDA-EMONDS

*Lehrstuhl für Tierzucht,
Technische Universität München,
Freising-Weihenstephan, Germany*



SPRINGER-SCIENCE+BUSINESS MEDIA, B.V.

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN 978-1-4020-2329-3 ISBN 978-1-4020-2330-9 (eBook)
DOI 10.1007/978-1-4020-2330-9

Printed on acid-free paper

All Rights Reserved
© 2004 Springer Science+Business Media Dordrecht
Originally published by Kluwer Academic Publishers in 2004
Softcover reprint of the hardcover 1st edition 2004

No part of the material protected by this copyright notice may be reproduced or
utilized in any form or by any means, electronic or mechanical,
including photocopying, recording or by any information storage and
retrieval system, without written permission from the copyright owner.

Contents

List of contributors	viii
Preface and acknowledgements	xii
Introduction	1
New uses for old phylogenies: an introduction to the volume <i>Olaf R. P. Bininda-Emonds</i>	3
1. Reviews of existing methods	15
The MRP method <i>Bernard R. Baum and Mark A. Ragan</i>	17
An assessment of matrix representation with compatibility in supertree construction <i>Howard A. Ross and Allen G. Rodrigo</i>	35
MRF supertrees <i>J. Gordon Burleigh, Oliver Eulenstein, David Fernández-Baca, and Michael J. Sanderson</i>	65
Everything you always wanted to know about average consensus, and more <i>François-Joseph Lapointe and Claudine Levasseur</i>	87
Tangled trees from multiple markers: reconciling conflict between phylogenies to build molecular supertrees <i>James Cotton and Roderic R. M. Page</i>	107

2. New supertree methods	127
Supertree methods for ancestral divergence dates and other applications <i>David Bryant, Charles Semple, and Mike Steel</i>	129
Supertree algorithms for nested taxa <i>Philip Daniel and Charles Semple</i>	151
Quartet supertrees <i>Raul Piaggio-Talice, J. Gordon Burleigh, and Oliver Eulensteiner</i>	173
Bayesian supertrees <i>Fredrik Ronquist, John Huelsenbeck, and Tom Britton</i>	193
3. Methodological considerations	225
Some desiderata for liberal supertrees <i>Mark Wilkinson, Joseph L. Thorley, Davide Pisani, François-Joseph Lapointe, and James O. McInerney</i>	227
Taxonomy, supertrees, and the Tree of Life <i>Roderic D. M. Page</i>	247
Garbage in, garbage out: data issues in supertree construction <i>Olaf R. P. Bininda-Emonds, Kate E. Jones, Samantha A. Price, Marcel Cardillo, Richard Grenyer and Andy Purvis</i>	267
Reconstructing divergence times for supertrees: a molecular approach <i>Rutger Vos and Arne Ø. Mooers</i>	281
Performance of supertree methods on various data set decompositions <i>Usman Roshan, Bernard M. E. Moret, Tiffani L. Williams, and Tandy Warnow</i>	301

4. A critical look at supertrees	329
Unrooted supertrees: limitations, traps, and phylogenetic patchworks <i>Sebastian Böcker</i>	331
The cladistics of matrix representation with parsimony analysis <i>Harold N. Bryant</i>	353
A critique of matrix representation with parsimony supertrees <i>John Gatesy and Mark S. Springer</i>	369
Supertrees, components and three-item data <i>David M. Williams</i>	389
5. Supertrees and their applications	409
A molecular supertree of the Artiodactyla <i>Annette S. Mahon</i>	411
Supertrees: using complete phylogenies in comparative biology <i>John L. Gittleman, Kate E. Jones, and Samantha A. Price</i>	439
Using supertrees to investigate species richness in grasses and flowering plants <i>Nicolas Salamin and T. Jonathan Davies</i>	461
Detecting diversification rate variation in supertrees <i>Brian R. Moore, Kai M. A. Chan, and Michael J. Donoghue</i>	487
Taxon index	535
Subject index	537

List of contributors

BERNARD R. BAUM

Eastern Cereal and Oilseed Research Centre,
Agriculture and Agri-Food Canada, Research
Branch, Neatby Building, 960 Carling Avenue,
Ottawa, ON, K1A 0C6, Canada

E-mail: baumbr@agr.gc.ca

OLAF R. P. BININDA-EMONDS

Lehrstuhl für Tierzucht, Technical University
of Munich, Alte Akademie 12, 85354
Freising–Weihenstephan, Germany

E-mail: Olaf.Bininda@tierzucht.tum.de

SEBASTIAN BÖCKER

AG Genominformatik, Technische Fakultät,
Universität Bielefeld, Postfach 100 131, 33501
Bielefeld, Germany

E-mail: boecker@CeBiTec.uni-bielefeld.de

TOM BRITTON

Department of Mathematics, Stockholm
University, SE-106 91, Stockholm, Sweden
E-mail: tomb@math.su.se

DAVID BRYANT

McGill Centre for Bioinformatics, Lyman Duff
Building, 3775 University Street, McGill
University, Montréal, QC, H3A 2B4, Canada
E-mail: bryant@mcb.mcgill.ca

HAROLD N. BRYANT

Royal Saskatchewan Museum, 2340 Albert
Street, Regina, SK, S4P 3V7, Canada

E-mail: HBryant@cyr.gov.sk.ca

J. GORDON BURLEIGH

Section of Ecology and Evolution, University
of California at Davis, Davis, CA, 95616, USA
E-mail: jgburleig@ucdavis.edu

MARCEL CARDILLO

Department of Biological Sciences, Imperial
College London, Silwood Park campus, Ascot
SL5 7PY, United Kingdom
E-mail: m.cardillo@imperial.ac.uk

KAI M. A. CHAN

Center for Conservation Biology, Department
of Biological Sciences, Stanford University,
Stanford, CA, 94305–5020, USA
E-mail: kaichan@stanford.edu

JAMES A. COTTON

Department of Zoology, The Natural History
Museum, Cromwell Road, London SW7 5BD,
United Kingdom
E-mail: james.cotton@nhm.ac.uk

PHILIP DANIEL

Biomathematics Research Centre, Department
of Mathematics and Statistics, University of
Canterbury, Christchurch, New Zealand
E-mail: pjd62@student.canterbury.ac.nz

T. JONATHAN DAVIES

Department of Biological Sciences, Imperial College London, Silwood Park campus, Ascot SL5 7PY, United Kingdom

E-mail: jon.davies@ic.ac.uk

MICHAEL J. DONOGHUE

Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT, 06520, USA

E-mail: michael.donoghue@yale.edu

OLIVER EULENSTEIN

Department of Computer Science, Iowa State University, Ames, IA, 50011–1040, USA

E-mail: oeulenst@cs.iastate.edu

DAVID FERNÁNDEZ-BACA

Department of Computer Science, Iowa State University, Ames, IA, 50011–1040, USA

E-mail: fernande@cs.iastate.edu

JOHN GATESY

Department of Biology, University of California at Riverside, Riverside, CA, 92521, USA

E-mail: johnga@ucr.edu

JOHN L. GITTLEMAN

Department of Biology, Gilmer Hall, University of Virginia, Charlottesville, VA, 22904–4328, USA

E-mail: JLGittleman@virginia.edu

RICHARD GRENYER

Department of Biology, Gilmer Hall, University of Virginia, Charlottesville, VA, 22904–4328, USA

E-mail: rich.grenyer@virginia.edu

JOHN P. HUELSENBECK

Section of Ecology, Behavior, and Evolution, Division of Biological Sciences, University of California at San Diego, San Diego, CA, 92093–0116, USA

E-mail: johnh@biomail.ucsd.edu

KATE E. JONES

Center for Environmental Research and Conservation, Columbia University, 1200 Amsterdam Avenue, MC 5556, New York, NY, 10027, USA

E-mail: kj2107@columbia.edu

FRANÇOIS-JOSEPH LAPointe

Département de sciences biologiques, Université de Montréal, C.P. 6128, Succursale Centre-ville, Montréal, QC, H3C 3J7, Canada

E-mail: Francois-Joseph.Lapointe@Umontreal.ca

CLAUDINE LEVASSEUR

Département de sciences biologiques, Université de Montréal, C.P. 6128, Succursale Centre-ville, Montréal, QC, H3C 3J7, Canada

E-mail: Claudine.Levasseur@Umontreal.ca

ANNETTE S. MAHON

University Museum of Zoology Cambridge, Department of Zoology, Downing Street, Cambridge CB2 3EJ, United Kingdom

E-mail: am354@hermes.cam.ac.uk

JAMES O. McINERNEY

Department of Biology, National University of Ireland, Maynooth, Co. Kildare, Ireland

E-mail: James.O.McInerney@may.ie

ARNE Ø. MOOERS

Department of Biological Sciences, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada

E-mail: amoers@sfu.ca

BRIAN R. MOORE

Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT, 06520, USA

E-mail: brian.moore@yale.edu

BERNARD M. E. MORET

Department of Computer Science, Farris Engineering Center #157, University of New Mexico, Albuquerque, NM, 87131–1386, USA
E-mail: moret@cs.unm.edu

RODERIC D. M. PAGE

Division of Environmental and Evolutionary Biology, Institute of Biomedical and Life Sciences, University of Glasgow, Glasgow G12 8QQ, United Kingdom
E-mail: r.page@bio.gla.ac.uk

RAUL PIAGGIO-TALICE

Department of Computer Science, Iowa State University, Ames, IA, 50011–1040, USA
E-mail: rpiaggio@iastate.edu

DAVIDE PISANI

Department of Zoology, The Natural History Museum, Cromwell Road, London SW7 5BD, United Kingdom
E-mail: D.Pisani@nhm.ac.uk

SAMANTHA A. PRICE

Department of Biology, Gilmer Hall, University of Virginia, Charlottesville, VA, 22904–4328, USA
E-mail: sp9b@virginia.edu

ANDY PURVIS

Department of Biological Sciences, Imperial College London, Silwood Park campus, Ascot SL5 7PY, United Kingdom

E-mail: a.purvis@imperial.ac.uk

MARK A. RAGAN

Institute for Molecular Bioscience, The University of Queensland, Brisbane, 4072 Queensland, Australia

E-mail: M.Ragan@imb.uq.edu.au

ALLEN G. RODRIGO

Bioinformatics Institute and the Allan Wilson Centre for Molecular Ecology and Evolution, School of Biological Sciences, University of Auckland, Private Bag 92019, Auckland, New Zealand

E-mail: a.rodrigo@auckland.ac.nz

FREDRIK RONQUIST

School of Computational Science and Information Technology, Florida State University, Tallahassee, FL, 32306–4120, USA

E-mail: ronquist@csit.fsu.edu

USMAN ROSHAN

Department of Computer Sciences, University of Texas at Austin, Austin, TX, 78712–1188, USA
E-mail: usman@cs.utexas.edu

HOWARD A. ROSS

Bioinformatics Institute and the Allan Wilson Centre for Molecular Ecology and Evolution, School of Biological Sciences, University of Auckland, Private Bag 92019, Auckland, New Zealand

E-mail: h.ross@auckland.ac.nz

NICOLAS SALAMIN

Felsenstein Lab, Department of Genome Sciences, Box 357730, University of Washington, Seattle, WA, 98195–7730, USA
E-mail: salamin@gs.washington.edu

MICHAEL J. SANDERSON

Section of Ecology and Evolution, University of California at Davis, Davis, CA, 95616, USA
E-mail: mjsanderson@ucdavis.edu

CHARLES SEMPLE

Biomathematics Research Centre, Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand
E-mail: c.semple@math.canterbury.ac.nz

MARK S. SPRINGER

Department of Biology, University of California at Riverside, Riverside, CA, 92521, USA
E-mail: mark.springer@ucr.edu

MIKE STEEL

Biomathematics Research Centre, Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand
E-mail: m.steel@math.canterbury.ac.nz

JOSEPH L. THORLEY

Fisheries Research Service, Freshwater Laboratory, Faskally, Pitlochry, Perthshire PH16 5LB, United Kingdom
E-mail: j.thorley@marlab.ac.uk

RUTGER A. VOS

Department of Biological Sciences, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada
E-mail: rvosa@sfu.ca

TANDY WARNOW

Department of Computer Sciences, University of Texas at Austin, Austin, TX, 78712–1188, USA
E-mail: tandy@cs.utexas.edu

MARK WILKINSON

Department of Zoology, The Natural History Museum, Cromwell Road, London SW7 5BD, United Kingdom
E-mail: marw@nhm.ac.uk

DAVID M. WILLIAMS

Department of Botany, The Natural History Museum, Cromwell Road, SW7 5BD London, United Kingdom
E-mail: dmw@nhm.ac.uk

TIFFANI L. WILLIAMS

Department of Computer Science, Farris Engineering Center #157, University of New Mexico, Albuquerque, NM, 87131–1386, USA
E-mail: tlw@cs.unm.edu

Preface and acknowledgements

The origins of this book lie in the Fifth Annual International Conference on Computational Molecular Biology (RECOMB 2001), held in Montréal, Canada from April 22–25, 2001. At this meeting, my Publishing Editor from Kluwer Academic Publishers, F. Robbert van Berckelaer, surveyed the participants for current “hot topics” in bioinformatics. Apparently, supertrees were mentioned often and my name in connection with them on more than one occasion. Upon returning to Dordrecht in the Netherlands, Robbert looked me up (it helped that I was only a few canals down the road in Leiden at the time) to see if I would be interested in producing a volume on supertrees. I agreed readily, feeling it to be an important topic and thinking that it all might be rather fun. Robbert and Kluwer showed great faith in the book from the beginning, and it is questionable whether this volume would exist at this time without their support. Thanks too to the anonymous RECOMBinites for suggesting my name in connection with a book on supertrees in the first place.

Much time has elapsed obviously since RECOMB 2001 and that initial meeting, such that this volume has spanned a pair of postdoctoral positions and supervisors. I thank Michael Richardson of Leiden University for his tremendous support and encouragement during the initial phases of the project, and Ruedi Fries of the Technische Universität München for allowing me to see it through to its completion. Financial support was provided by both the van der Leeuw fonds (Leiden) and the BMBF through the “Bioinformatics for the Functional Analysis of Mammalian Genomes” (BFAM) project (Freising).

I am grateful to John Gittleman, Julie Lockwood, and Rod Page for their advice regarding some of the practical issues in putting together an edited volume. John and Rod, in particular, answered my many, many queries with the utmost patience and helpfulness.

Naturally, the bulk of the effort in creating this book lies with the chapter authors. I thank them for making my job as easy as it probably could be, and especially for sharing my enthusiasm for the project. On the latter count, I would note that only about half of the chapters were “invited” on my part;

the other half invited themselves as word went around that a book on supertrees was taking shape. One of the latter, self-invited chapters was even written in response to a problem raised in one of the former chapters! Altogether, the various authors have helped make this volume a much broader, diverse, and complete representation of the field of supertree research than I ever could have written single-handedly (which was one of the initial suggestions).

Writing, of course, is only half the story. The often-untold other half is the reviewing, and all contributions in this volume were peer-reviewed. For their help in this thankless task and for doing it so punctually, I thank Paul-Michael Agapow, Bernard Baum, Vincent Berry, David Bryant, Harold Bryant, Gordon Burleigh, Sebastian Böcker, Marcel Cardillo, Mike Charleston, Guy Cucumel, Bill Day, Kevin de Queiroz, Mike Dodd, Olivier Gascuel, John Gatesy, Pablo Goloboff, Colin Groves, Katharina Huber, Volker Hösel, Kate Jones, François-Joseph Lapointe, Peter Mayhew, Arne Mooers, Vincent Moulton, Rod Page, Davide Pisani, Samantha Price, Andy Purvis, Mark Ragan, Allen Rodrigo, Fredrik Ronquist, Nicolas Salamin, Charles Semple, Mike Steel, Rutger Vos, Mark Wilkinson, David Williams, Tiffani Williams, Stephen Willson, Elizabeth Zimmer, and several anonymous reviewers. Their efforts in the way of comments, corrections, and suggestions have helped to ensure the accuracy and quality of all the contributed papers.

Finally, I thank my wife, Julia Gockel, for her unending support, understanding, and patience for the enduring project that was *The Book*. At times, she worked harder than I did to free up time for me to work on it, for which I cannot thank her enough.

And, yeah, it all was kind of fun!

Freising, Germany

Olaf R. P. Bininda-Emonds

Introduction

NEW USES FOR OLD PHYLOGENIES

An introduction to the volume

Olaf R. P. Bininda-Emonds

1. Introduction

“This is a paper with an attitude problem. This may sound facetious, but is meant in all seriousness. It has in my opinion entirely the wrong attitude to phylogenetic reconstruction and indeed to the entire scientific process.”

From an anonymous review of the carnivore supertree of Bininda-Emonds *et al.* (1999)

What are supertrees and what is all the fuss about?

These are two of the questions that this volume will attempt to answer. A brief answer to the former is that supertree construction is a phylogenetic approach that combines tree topologies instead of the primary character data that they are based on. It differs from traditional consensus techniques, which also combine tree topologies, in that the constituent (or “source”) trees need only be overlapping, and not identical, with respect to the terminal taxa they contain. As such, the resulting supertree can be, and usually is, larger than any of the source trees contributing to it. Supertrees thus represent an exciting opportunity to build more comprehensive phylogenies: in essence, new uses for old phylogenies (with apologies to Harvey *et al.*, 1996). However, the use of tree topologies and not primary character data as the source data has attracted much criticism (e.g., Rodrigo, 1993; Slowinski and Page, 1999; Novacek, 2001; Springer and de Jong, 2001; Gatesy *et al.*,

Bininda-Emonds, O. R. P. (ed.) Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life, pp. 3–14. Computational Biology, volume 3 (Dress, A., series ed.).

© 2004 Kluwer Academic Publishers.

2002), such that supertree construction is an increasingly popular, but highly controversial approach in phylogenetic systematics.

Although supertree construction has attracted increasing attention only in the past few years, the fundamental idea behind it — that of combining numerous source trees to yield a single, more inclusive tree — has a longer, if unrecognized, history. The process of synthesizing systematic knowledge by cutting and pasting together evolutionary trees as “informal” supertrees is probably nearly as old as systematics itself. Even today, any detailed depiction of the single Tree of Life (e.g., the Tree of Life Web Project; <http://tolweb.org/tree/phylogeny.html>), if not any conception that we have of it, can be achieved only using supertrees; the largest phylogenies based on primary character data are on the order of thousands of species only (e.g., Källersjö *et al.*, 1998; Johnson, 2001). Although informal supertrees continue to be constructed (e.g., Garland *et al.*, 1993; Kennedy *et al.*, 1996; Ortolani, 1999; Webb, 2000; Cardillo and Bromham, 2001; Hall and Harvey, 2002), this volume deals exclusively with the more formal supertree construction techniques.

The formalization of supertree methodology, and the term supertree itself, stem from Allan Gordon’s (1986) seminal paper. In this paper, Gordon described the supertree equivalent of strict consensus, whereby the supertree contained only those groups found on or implied jointly by all the source trees. However, it did not have much of an immediate impact for several reasons. First, the paper was published in a mathematical journal, whereas the current popularity of supertrees arguably derives largely from the biological community. Second, the current interest in supertrees derives in large measure from their ability to build very large phylogenies of hundreds of species, something that has become computationally feasible within only the past decade at best. Finally, Gordon’s method was limited to overlapping source trees that were compatible: they could differ from one another, but not actually conflict. Thus, the method was of limited utility. As most systematists know, phylogenies usually conflict with one another.

The breakthrough for supertrees came in 1992, when Bernard Baum, Jeff Doyle, and Mark Ragan independently described the supertree technique known as simply MRP (matrix representation with parsimony; Baum, 1992; Doyle, 1992; Ragan, 1992). MRP, like Brook’s Parsimony Analysis (Brooks, 1981), makes use of additive binary coding (Farris *et al.*, 1970) to represent a given tree in matrix format. The “matrix representations” of the different source trees are then combined into a single matrix that can be analyzed using any desired optimization criterion (but usually parsimony). This procedure removed the fundamental limitation of Gordon’s strict supertree method: all overlapping source trees could now be combined as a supertree, regardless of how much they conflicted with one another. At the

same time, these trees could be derived from all possible data types (including no data whatsoever!), overcoming the limitation of combined-data (“total evidence” or “supermatrix”) approaches (*sensu* Kluge, 1989; Sanderson *et al.*, 1998; respectively) that the data types be analyzable using a single optimization criterion.

It was Andy Purvis who perhaps first realized the tremendous potential of supertrees to biology. His MRP supertree of all 203 extant species of primate represented the first, complete (super)tree of a significant clade that was based on an objective methodology (Purvis, 1995a). Its large size, its unprecedented completeness, and its high amount of resolution demonstrated what supertree construction could achieve. Moreover, Purvis showed immediately that supertrees have biological utility beyond their obvious systematic value when he used his primate supertree to answer numerous macroevolutionary questions in a phylogenetic framework (Purvis *et al.*, 1995). The primate supertree has gone on to become perhaps the reference standard for supertrees. It has been updated twice (Purvis and Webster, 1999; Vos and Mooers, in prep.), and is often the tree against which new methodologies are tested (e.g., Moore *et al.*, 2004; Vos and Mooers, 2004).

However, in some ways, Purvis caught the phylogenetic community unawares. The next major supertree, that of the mammalian order Carnivora that I published (Bininda-Emonds *et al.*, 1999), took another four years to be published, largely as a result of hostile reviews (see above). An evolutionary journal thought that the carnivore supertree was “too taxonomic”. A taxonomic journal thought the reverse: it was “too evolutionary”. The supertree was published eventually in a review journal, despite arguably containing very little explicit review material. Many other supertrees studies have faced equally difficult routes to publication and critiques of the supertree approach are appearing more frequently (see above).

Today, supertree construction is an active field of theoretical, practical, and applied research in mathematics, algorithmics, computer sciences, and biology. This multidisciplinary, bioinformatic nature to supertree research has produced numerous advances and developments in a short time: supertrees are being constructed at an increasing pace, and supertree methods continue to be developed and improved. Supertree construction is also mentioned increasingly as perhaps being a key element in our efforts to reconstruct the Tree of Life (e.g., Soltis and Soltis, 2001; Pennisi, 2003). It has certainly illuminated some of the largest portions of the Tree to date, and will continue to do so for some time to come. However, in so doing, the role of supertrees might change from simply combining existing information to being an important analytical tool to search large character matrices efficiently (Bininda-Emonds *et al.*, 2002). Supertrees, like the species they depict, are also continuing to evolve.

2. Structure of the volume

This volume is divided roughly, if not somewhat arbitrarily, into the following five sections: existing supertree methods, new supertree methods, methodological considerations, a critical look at supertrees, and supertrees and their applications. Here, I examine each of these sections, and introduce briefly the chapters within them, in turn.

2.1 Existing supertree methods

Together with the following section, this is the largest component of the book, reflecting the plethora of supertree methods that have been and continue to be developed. More than a dozen major methods, and numerous variants on these methods, exist currently (Table 1). This section provides reviews of most of the current methods. Only reviews of Gordon's strict supertrees (Gordon, 1986), semi-strict supertrees (Goloboff and Pol, 2002), and (modified) MINCUTSUPERTREE (Semple and Steel, 2000; Page, 2002) are lacking, although the latter does make an appearance in several chapters in the book.

The first chapter, perhaps fittingly, is about MRP, by far the most popular of the many supertrees techniques. In this chapter, Bernard Baum and Mark Ragan discuss their motivations in developing MRP, review several MRP-related issues, and argue strongly for a continuing role of MRP in large-scale phylogeny reconstruction. But, as Baum and Ragan (1993) themselves have pointed out, parsimony is not the only option for analyzing representations of source trees.

The two following chapters review alternative methods for analyzing matrix representations of source trees. First, Howard Ross and Allen Rodrigo explore an idea first raised by Purvis (1995b) nearly a decade ago (and followed up by Rodrigo, 1996; Pisani, 2002): using compatibility instead of parsimony to analyze the matrix representations of the source trees. Gordon Burleigh and colleagues then review the concept of (minimum) flip supertrees, in which analysis proceeds not by optimizing the combined matrix representations, but by altering ("flipping") individual cells in the matrix so as to remove any conflict between them.

The section concludes with two chapters about long-existing, but perhaps unappreciated supertree methods: the average consensus and gene tree parsimony. François-Joseph Lapointe and Claudine Levasseur discuss the natural extension of the average consensus (Lapointe and Cucumel, 1997) to the supertree setting. This method is notable because it uses an alternative form of matrix representation based on path-length distance matrices rather than an MRP-like membership criterion. As such, it can maintain and utilize

Table 1. The major supertree methods and their variants. The methods are subdivided according to whether they produce a supertree that either summarizes common structure among the source tree (“agreement supertrees”) or maximizes the fit to the set of source trees according to some objective function (“optimization supertrees”). Methods in bold face are either reviewed or introduced in this volume.

Agreement supertrees	Optimization supertrees
Gordon’s strict	Average consensus (also known as matrix representation with distances, MRD)
MINCUTSUPERTREE, including: modified MinCutSupertree	Bayesian supertrees
RANKEDTREE	Gene tree parsimony
SEMI-LABELLED- and ANCESTRALBUILD	Matrix representation with compatibility (MRC)
Semi-strict	Matrix representation with flipping (MRF; also known as MinFlip supertrees)
Strict consensus merger	Matrix representation with parsimony (MRP), including: Purvis sister-group coding Irreversible MRP
	Quartet supertrees

branch-length information in the source trees unlike most supertree methods. Likewise, the extension to the supertree setting of gene tree parsimony — a technique developed originally to reconcile conflicting phylogenies in co-evolutionary, host-parasite, biogeography, or gene family evolution studies (Slowinski and Page, 1999) — is described by James Cotton and Rod Page.

2.2 New supertree methods

An active area in supertrees is the continuing development of different methods. Of the methods listed in Table 1, the majority date from the past few years only. This includes four new methods described for the first time in this volume.

Three existing supertree methods — Gordon’s strict method, MINCUTSUPERTREE, and modified MINCUTSUPERTREE — all rely on the BUILD algorithm of Aho *et al.* (1981), which, interestingly, was developed for other purposes entirely (namely, relational databases!) and actually predates the formalization of the supertree approach. Two of the new methods in this section derive from further modifications of BUILD. First, David Bryant and colleagues describe RANKEDTREE, an algorithm that is able to incorporate both relative and absolute dating information from the source trees so as to directly produce a supertree with divergence-date

estimates. Then, Philip Daniel and Charles Semple provide a solution to a problem raised by Rod Page in a chapter that appears later in the book, and introduce modifications to BUILD that allow sets of source trees with nested terminal taxa to be combined. Their algorithms make use of the fact that not only the terminals, but also the nodes in a phylogeny might be labeled.

Although the idea of supertree methods based on quartets has been raised previously (e.g., Thorley and Page, 2000; Pisani and Wilkinson, 2002), Raul Piaggio-Talice and colleagues provide one of the first working descriptions of a quartet-based supertree method. In their chapter, they explore the performance of two quartet-supertree methods that they based on the character-based quartet methods of Stephen Willson (1999, 2001). The final chapter by Fredrik Ronquist *et al.* continues the expansion of Bayesian methodology into evolutionary biology by introducing Bayesian supertrees. Interestingly, the heart of Bayesian supertrees is the same matrix representation of supertrees used by MRP and several other related methods. As it turns out, the matrix representations provide an excellent summary of the structure of a tree (in the form of taxon bipartitions) that translates well to the Bayesian framework.

2.3 Methodological considerations

In supertree construction, as in conventional phylogenetics, there are always questions about how to apply or expand upon those methods that do exist. Some of these issues are dealt with directly when developing a supertree method, but many more general ones still exist. This section examines a handful of some of these many issues.

The previous two sections in the book and Table 1 together indicate that supertree methods, each with slightly different properties, abound. But, what properties should a (good) supertree method have? Taking their cue from the axiomatic approach that is common in the mathematical literature (e.g., McMorris and Neumann, 1983; Barthélemy *et al.*, 1995; Steel *et al.*, 2000), Mark Wilkinson and colleagues propose a list of desirable features (“desiderata”) that are based on the goals of accuracy and practicality. They then attempt to characterize the many liberal (or optimization) supertree techniques for these desiderata.

There then follow two chapters that look more closely at the raw data of a supertree analysis, the source trees themselves. In the first chapter, Rod Page poses numerous questions and challenges for supertree researchers that derive from taxonomic considerations. As mentioned, one problem was solved subsequently by Daniel and Semple (but appears in a previous chapter). Page also presents several answers himself, in particular using the concept of a classification graph to potentially increase the degree of

taxonomic overlap between source trees. Then, together with several members of the “Mammal SuperTeam”, I outline the protocol for source-tree collection and manipulation that we established as part of our efforts to construct a supertree of most extant mammalian species. We hope that our protocol, suitably adjusted for the supertree project in question, will help other researchers in constructing their supertrees.

An active area of research in conventional, data-based phylogenetics is the development of methods to infer divergence times for phylogenies from DNA sequence information (e.g., Rambaut and Bromham, 1998; Thorne *et al.*, 1998; Huelsenbeck *et al.*, 2000a; Yoder and Yang, 2000; Sanderson, 2002; Thorne and Kishino, 2002). Previous efforts in this area for supertrees have been less formal (e.g., Purvis, 1995a; Bininda-Emonds *et al.*, 1999). Rutger Vos and Arne Mooers address this deficiency and describe a method to fit available DNA sequence data to a supertree topology. Using their method of “gene shopping” and “taxon shopping”, they infer divergence dates for a new update to the primate supertree, and compare their estimates to those from Purvis’s (1995a) original supertree of the Primates.

Usman Roshan and colleagues then provide the exception to the title of this introduction (if not the book), and describe what might well be the future of supertree research. Rather than use supertrees to piece together previously existing information, Roshan *et al.* explore the idea of supertrees as part of a divide-and-conquer strategy, in which a large data matrix is broken down and analyzed as easier subproblems, which are then recombined using supertrees to give the global answer. A similar strategy is provided also by Gordon Burleigh and colleagues in their earlier chapter on MRF supertrees, in which their biclique approach is used to decompose a large, incomplete data matrix into smaller, complete submatrices, the results of which can be combined with supertrees.

2.4 A critical look at supertrees

As I indicated above, supertrees are not an uncontroversial area of research. Interestingly, however, the harshest criticisms of the supertree approach originate almost exclusively from the phylogenetic systematics community (see above), a group for which supertrees might be thought to hold the greatest immediate benefit. Perhaps connected with this fact, it remains that few major supertree analyses have been published in and of themselves in one of the leading systematic or taxonomic journals. The exception is the genus-level grass supertree of Salamin *et al.* (2002), which was published in *Systematic Biology*, but had a strong methodological component to it. This section casts a critical eye at supertrees, indicating perhaps that there is still room for improvement in this young field.

The first chapter in this section is not so much a critique of the supertree approach, but rather serves to point out some fundamental limitations when attempting to build unrooted supertrees. These problems are perhaps underappreciated by biologists, who tend to work with rooted trees; building rooted supertrees, as it turns out, faces far fewer such problems. In his chapter, Sebastian Böcker reviews some of these known limitations, but also reveals instances under which it is possible to build an unrooted supertree efficiently.

By virtue of being the most popular supertree method, MRP is also the one that has attracted the most direct criticism. (In fact, many of the alternative supertree methods have been developed to address perceived shortcomings with MRP.) The remaining chapters in this section continue this critical examination of MRP, each from a slightly different perspective.

In the biological systematics community, parsimony is linked intimately with cladistics. Consequently, there is a natural tendency to associate MRP and cladistics, and several critiques of MRP derive from attempting to interpret the method and its results in a cladistic framework. Harold Bryant considers this relationship in more detail, and examines explicitly how well MRP meets the assumptions of cladistic analysis.

Supertree construction is but one approach to combine existing phylogenetic information to derive more comprehensive phylogenies. Another is the direct combination of the primary character data, known variously as the supermatrix (*sensu* Sanderson *et al.*, 1998) or total evidence approach (*sensu* Kluge, 1989). In their chapter, John Gatesy and Mark Springer contrast these two approaches to expand on their previous criticisms of both MRP and the supertree approach as a whole (Springer and de Jong, 2001; Gatesy *et al.*, 2002).

Finally, David Williams examines the representation of source tree topologies, not as binary characters as is usually done, but as three-item statements encoding relationships within the tree. Intriguingly, he suggests that any undesirable aspects to MRP might derive not only from its use of what he holds to be an inferior form of matrix representation (i.e., binary coding), but, more importantly, from a fundamental shortcoming in parsimony optimization itself.

2.5 Supertrees and their applications

The final section departs from the methodological focus of the previous chapters to examine (biological) applications of supertrees. In recent years, the importance of examining biological questions in a phylogenetic framework has become increasingly appreciated. As phylogenetic hypotheses, supertrees can be used like any other conventionally derived

phylogeny. However, the ability of supertree construction to yield large complete phylogenetic estimates for a given clade allows biologists to potentially examine questions of greater scope and with more statistical power than would be possible with conventionally derived phylogenies.

The section begins with a large-scale supertree, that of 171 species of mammalian order Artiodactyla by Annette Mahon. Although the supertree is not complete — it is missing the Cetacea (now agreed widely to cluster within Artiodactyla) and several artiodactyl species — this was a result of conscious decisions on her part. This illustrates that supertrees need not be complete, but, like any phylogeny, can be tailor-made to suit a given objective and is also subject to the same issues of data quality and availability.

John Gittleman and colleagues then provide a general, wide-ranging review of how supertrees as large, complete phylogenetic estimates have been and perhaps could be used in biological research beyond their obvious utility for descriptive systematics. Although bigger might be better, they strike a note of caution as well, and point out limitations to supertrees that might occasionally make them inappropriate as a basis upon which to draw biological inferences.

The final two chapters in the volume use supertrees to examine macroevolutionary patterns in three diverse clades (grasses, angiosperms, and primates). All three analyses are possible because of the high level of taxonomic completeness that is potentially achievable through supertree construction. First, Nicolas Salamin and Jon Davies test numerous key-innovation hypotheses from the literature that attempt to relate species richness in grasses and angiosperms with certain morphological adaptations. Finally, Brian Moore and colleagues expand on their previous work with topology-based methods to investigate diversification rates (e.g., Chan and Moore, 2002), and develop a suite of shift statistics that can pinpoint where significant changes in rate have occurred on a supertree. They then apply their statistics to the dated primate supertree of Purvis (1995a), and compare their whole-tree macroevolutionary inferences to those of Purvis *et al.* (1995), which were derived using statistics that require divergence date estimates.

3. A last word

This volume hopefully reflects the current diversity of supertree research, with contributions from the different communities of mathematics, algorithmics, computer sciences, and biology. It is important to note that the writing conventions and even the “languages” of these communities differ

greatly (particularly between the “extremes” of mathematics and biology). I have made a conscious decision to retain these differences, so as to maximize the impact and accuracy of each chapter for their respective community. As such, the appeal and accessibility of the different chapters will undoubtedly vary greatly depending on the background of the reader. However, in the end, I hope that the book serves both to provide an introduction to supertree construction and to highlight current research areas and issues.

Acknowledgements

Financial support from the BMBF, through the “Bioinformatics for the Functional Analysis of Mammalian Genomes” (BFAM) project is gratefully acknowledged.

References

- AHO, A. V., SAGIV, Y., SZYMANSKI, T. G., AND ULLMAN, J. D. 1981. Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. *SIAM Journal on Computing* 10:405–421.
- BARTHÉLEMY, J.-P., McMORRIS, F. R., AND POWERS, R. C. 1995. Stability conditions for consensus functions defined on n -trees. *Mathematical Computer Modeling* 22:79–87.
- BAUM, B. R. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon* 41:3–10.
- BAUM, B. R. AND RAGAN, M. A. 1993. Reply to A.G. Rodrigo’s “A comment on Baum’s method for combining phylogenetic trees”. *Taxon* 42:637–640.
- BININDA-EMONDS, O. R. P., GITTLEMAN, J. L., AND PURVIS, A. 1999. Building large trees by combining phylogenetic information: a complete phylogeny of the extant Carnivora (Mammalia). *Biological Reviews* 74:143–175.
- BININDA-EMONDS, O. R. P., GITTLEMAN, J. L., AND STEEL, M. A. 2002. The (super)tree of life: procedures, problems, and prospects. *Annual Review of Ecology and Systematics* 33:265–289.
- BROOKS, D. R. 1981. Hennig’s parasitological method: a proposed solution. *Systematic Zoology* 30:229–249.
- CARDILLO, M. AND BROMHAM, L. 2001. Body size and risk of extinction in Australian mammals. *Conservation Biology* 15:1435–1440.
- CHAN, K. M. A. AND MOORE, B. R. 2002. Whole-tree methods for detecting differential diversification rates. *Systematic Biology* 51:855–865.
- DOYLE, J. J. 1992. Gene trees and species trees: molecular systematics as one-character taxonomy. *Systematic Botany* 17:144–163.
- FARRIS, J. S., KLUGE, A. G., AND ECKHARDT, M. J. 1970. A numerical approach to phylogenetic systematics. *Systematic Zoology* 19:172–191.
- GARLAND, T., JR, DICKERMAN, A. W., JANIS, C. M., AND JONES, J. A. 1993. Phylogenetic analysis of covariance by computer simulation. *Systematic Biology* 42:265–292.

- GATESY, J., MATTHEE, C., DESALLE, R., AND HAYASHI, C. 2002. Resolution of a supertree / supermatrix paradox. *Systematic Biology* 51:652–664.
- GOLOBOFF, P. A. AND POL, D. 2002. Semi-strict supertrees. *Cladistics* 18:514–525.
- GORDON, A. D. 1986. Consensus supertrees: the synthesis of rooted trees containing overlapping sets of labeled leaves. *Journal of Classification* 3:31–39.
- HALL, J. P. W. AND HARVEY, D. J. 2002. Basal subtribes of the Nymphidiini (Lepidoptera: Riodinidae): phylogeny and myrmecophily. *Cladistics* 18:539–569.
- HARVEY, P. H., LEIGH BROWN, A. J., MAYNARD SMITH, J., AND NEE, S. (eds) 1996. *New Uses for New Phylogenies*. Oxford University Press, Oxford.
- HUELSENBECK, J. P., LARGET, B., AND SWOFFORD, D. 2000. A compound Poisson process for relaxing the molecular clock. *Genetics* 154:1879–1892.
- JOHNSON, K. P. 2001. Taxon sampling and the phylogenetic position of Passeriformes: evidence from 916 avian cytochrome *b* sequences. *Systematic Biology* 50:128–136.
- KÄLLERSJÖ, M., FARRIS, J. S., CHASE, M. W., BREMER, B., FAY, M. F., HUMPHRIES, C. J., PETERSEN, G., SEBERG, O., AND BREMER, K. 1998. Simultaneous parsimony jackknife analysis of 2538 *rbcL* DNA sequences reveals support for major clades of green plants, land plants, seed plants and flowering plants. *Plant Systematics and Evolution* 213:259–287.
- KENNEDY, M., SPENCER, H. G., AND GRAY, R. D. 1996. Hop, step and gape: do the social displays of the Pelecaniformes reflect their phylogeny? *Animal Behaviour* 51:273–291.
- KLUGE, A. G. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). *Systematic Zoology* 38:7–25.
- LAPOINTE, F.-J. AND CUCUMEL, G. 1997. The average consensus procedure: combination of weighted trees containing identical or overlapping sets of taxa. *Systematic Biology* 46:306–312.
- MCMORRIS, F. R. AND NEUMANN, D. 1983. Consensus functions defined on trees. *Mathematical Social Sciences* 4:131–136.
- MOORE, B. R., CHAN, K. M. A., AND DONOGHUE, M. J. 2004. Detecting diversification rate variation in supertrees. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 487–533. Kluwer Academic, Dordrecht, the Netherlands.
- NOVACEK, M. J. 2001. Mammalian phylogeny: genes and supertrees. *Current Biology* 11:R573–R575.
- ORTOLANI, A. 1999. Spots, stripes, tail tips and dark eyes: predicting the function of carnivore colour patterns using the comparative method. *Biological Journal of the Linnean Society* 67:433–476.
- PAGE, R. D. M. 2002. Modified mincut supertrees. In R. Guigó and D. Gusfield (eds), *Algorithms in Bioinformatics, Second International Workshop, WABI 2002, Rome, Italy, September 17–21, 2002, Proceedings*, pp. 537–552. Springer, Berlin.
- PENNISI, E. 2003. Modernizing the Tree of Life. *Science* 300:1692–1697.
- PISANI, D. 2002. *Comparing and Combining Data and Trees in Phylogenetic Analysis*. Ph.D. dissertation, Department of Earth Sciences, University of Bristol, United Kingdom.
- PISANI, D. AND WILKINSON, M. 2002. Matrix representation with parsimony, taxonomic congruence, and total evidence. *Systematic Biology* 51:151–155.
- PURVIS, A. 1995a. A composite estimate of primate phylogeny. *Philosophical Transactions of the Royal Society of London B* 348:405–421.
- PURVIS, A. 1995b. A modification to Baum and Ragan's method for combining phylogenetic trees. *Systematic Biology* 44:251–255.

- PURVIS, A., NEE, S., AND HARVEY, P. H. 1995. Macroevolutionary inferences from primate phylogeny. *Proceedings of the Royal Society of London B* 260:329–333.
- PURVIS, A. AND WEBSTER, A. J. 1999. Phylogenetically independent comparisons and primate phylogeny. In P. C. Lee (ed.), *Comparative Primate Socioecology*, pp. 44–70. Cambridge University Press, Cambridge.
- RAGAN, M. A. 1992. Phylogenetic inference based on matrix representation of trees. *Molecular Phylogenetics and Evolution* 1:53–58.
- RAMBAUT, A. AND BROMHAM, L. 1998. Estimating divergence dates from molecular sequences. *Molecular Biology and Evolution* 15:442–448.
- RODRIGO, A. G. 1996. On combining cladograms. *Taxon* 45:267–274.
- SALAMIN, N., HODKINSON, T. R., AND SAVOLAINEN, V. 2002. Building supertrees: an empirical assessment using the grass family (Poaceae). *Systematic Biology* 51:136–150.
- SANDERSON, M. J. 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Molecular Biology and Evolution* 19:101–109.
- SANDERSON, M. J., PURVIS, A., AND HENZE, C. 1998. Phylogenetic supertrees: assembling the trees of life. *Trends in Ecology and Evolution* 13:105–109.
- SEMPLE, C. AND STEEL, M. 2000. A supertree method for rooted trees. *Discrete Applied Mathematics* 105:147–158.
- SLOWINSKI, J. B. AND PAGE, R. D. M. 1999. How should species phylogenies be inferred from sequence data? *Systematic Biology* 48:814–825.
- SOLTIS, P. S. AND SOLTIS, D. E. 2001. Molecular systematics: assembling and using the Tree of Life. *Taxon* 50:663–677.
- SPRINGER, M. S. AND DE JONG, W. W. 2001. Phylogenetics. Which mammalian supertree to bark up? *Science* 291:1709–1711.
- STEEL, M., DRESS, A. W. M., AND BÖCKER, S. 2000. Simple but fundamental limitations on supertree and consensus tree methods. *Systematic Biology* 49:363–368.
- THORLEY, J. L. AND PAGE, R. D. 2000. RadCon: phylogenetic tree comparison and consensus. *Bioinformatics* 16:486–7.
- THORNE, J. L., KISHINO, H., AND PAINTER, I. S. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution* 15:1647–1657.
- THORNE, J. L. AND KISHINO, H. 2002. Divergence time and evolutionary rate estimation with multilocus data. *Systematic Biology* 51:689–702.
- VOS, R. A. AND MOOERS, A. Ø. 2004. Reconstructing divergence times for supertrees. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 281–299. Kluwer Academic, Dordrecht, the Netherlands.
- WEBB, C. O. 2000. Exploring the phylogenetic structure of ecological communities: an example for rain forest trees. *American Naturalist* 156:145–155.
- WILLSON, S. J. 1999. Building phylogenetic trees from quartets by using local inconsistency measures. *Molecular Biology and Evolution* 16:685–693.
- WILLSON, S. J. 2001. An error correcting map for quartets can improve the signals for phylogenetic trees. *Molecular Biology and Evolution* 18:344–351.
- YODER, A. D. AND YANG, Z. H. 2000. Estimation of speciation dates using local molecular clocks. *Molecular Biology and Evolution* 17:1081–1090.

1. Reviews of existing methods

Chapter 1

THE MRP METHOD

Bernard R. Baum and Mark A. Ragan

Abstract: Matrix representation with parsimony (MRP) is a supertree approach for analyzing multiple hierarchical trees, and through them multiple data sets, within a unified phylogenetic inference. Unlike consensus methods, it is based on nodes (subtrees) and not on full trees. This makes it possible to draw data sets with different but overlapping phyletic coverage into a common analysis. Our original method has provided a platform for subsequent modifications with respect to coding, weighting, transformations, and resolution of ambiguities and conflicts. Further extensions can be envisioned to improve not only performance in unified phylogenetic inference from large (e.g., genomic) and/or heterogeneous data sets, but also in the quantitative comparison of trees and subtrees.

Keywords: gene trees; genomics; matrix representation with parsimony; phylogenetics; supertrees

1. Introduction

In this chapter we explain the motivation and conceptual basis of the MRP method, discuss its benefits and limitations, indicate how MRP has been applied to biological problems, and consider future uses and extensions. We also comment briefly on some mathematical and algorithmic properties of MRP.

The method now called matrix representation with parsimony (MRP) was first presented at the Fourth International Congress of Systematic and Evolutionary Biology (Baum, 1990). Essentially the same method was described at the Fifth Annual Meeting of the Canadian Institute for Advanced Research, Program in Evolutionary Biology (Ragan, 1991). We

described the method in more detail independently and published the first MRP supertrees the following year (Baum, 1992; Ragan, 1992a,b). Soon thereafter, Rodrigo (1993) criticized the MRP method in comments on Baum's (1992) paper. By then aware of each other's work, we published a rejoinder to Rodrigo in the same issue (Baum and Ragan, 1993).

Although we developed the method independently, we were motivated by similar considerations. First, we were, and remain, interested in phylogenetic relationships among organisms (species), and believed that we could improve the quality of species trees by integrating results from multiple individual data sets, including gene and protein families. In those pre-genomic years, it was assumed that all genomic regions have the same history — a history that might, however, be obscured, either because individual genes or proteins contain too little information or because specific genes might be situated in atypical regions of the genome (e.g., regions of unusual G + C content). Genes (and the corresponding proteins) were seen as samples of genomes; by sampling multiple genes, we should increase the information content of our data and guard against region-specific bias. However, there were practical problems to overcome. Some genomic regions had been sampled at the DNA level, others by protein sequencing. Memory limitations in desktop computers and preset data lengths in popular phylogenetic software made it difficult to analyze large primary data sets. Many trees were appearing in the literature or were available from colleagues — could we not somehow use them directly?

Second, we wanted to synthesize molecular-sequence with non-sequence data. There are interesting historical and sociological issues around the extent to which the “molecular Tree of Life” paradigm had supplanted organism-based concepts of biodiversity and phylogenetics by 1990, but both of us were still working with morphologically or physiologically based phenotypic data in addition to molecular sequences. More broadly, we recognized a need to be able to mobilize discrete-state character data (e.g., presence/absence of morphological features), pairwise distances (e.g., serological data), and protein and nucleotide sequences into a joint analysis that might improve the approximation to the species tree. Incompatible data types (i.e., those based on different models, or those requiring different analytical methods or optimization criteria) cannot be analyzed jointly using the total evidence (TE) approach of Kluge (1989), in which data matrices are simply adjoined before analysis (e.g., by parsimony). Neither models nor software exist to analyze TE matrices containing multiple types of data (e.g., nucleotides, amino acids, pairwise distances and morphological characters).

Third, we had to work with data sets that shared only a portion of their evolutionary units in common. Even the best-studied organisms were still largely *terra incognita* at the molecular-sequence level, and few large

molecular-sequence data sets covered exactly the same set of organisms as a result. Little did we suspect that this situation was not merely a sampling artifact; only with the rise of microbial genomics has it begun to be appreciated how diverse genomes actually are with respect to gene content! This situation ruled out consensus as a useful tool in this application because consensus methods require that all component trees either share the same leaves, or (following Gordon, 1986) are perfectly compatible.

Finally, we wanted our inferences based on multiple data sets to reflect the relative strengths or values of those data sets. This happens naturally when individual data sets of the same data type (e.g., protein sequences) are combined (i.e., adjoined); but, as noted above, hardware and software limitations made it impossible for us to combine even primary sequence data, much less other data types. It did not escape our notice that a matrix-based approach should make it possible, at least in principle, to weight source data sets differentially (based, for example, on relative gene lengths or numbers of informative sites). It is fair comment that MRP as typically applied today does not achieve this goal fully, but the method does allow for weighting in a way that consensus methods do not.

2. The MRP method

The definitive feature of MRP is that it is based on the combined analysis of a set of trees, not on the direct combination of data underlying these trees as in TE. Each tree (component or source tree) in the set is viewed as a hierarchically ordered collection of nodes. Information about these nodes is recoded and combined into a single matrix, which is then analyzed to produce a synthesis of topological information over all source trees.

More precisely, our algorithm consists of five steps:

1. Given a set of trees inferred from different data sets, select all trees to be combined.
2. If some or all trees are unrooted, root all the trees with the same taxon. If some trees are rooted already with a different taxon, then re-root. Alternatively, or if no taxon is common to all trees, root all trees with a dummy (all-zero) outgroup. At this stage, internal nodes can be labeled optionally to assist in step 3 (see Figure 1).
3. Express the topology of each tree by additive binary coding, following Farris *et al.* (1970). For step-by-step examples, see Brooks (1981; also Figure 1).
4. Adjoin the resulting binary-coded matrices to form a single matrix of binary elements. Code the missing taxa (leaves) with a missing data

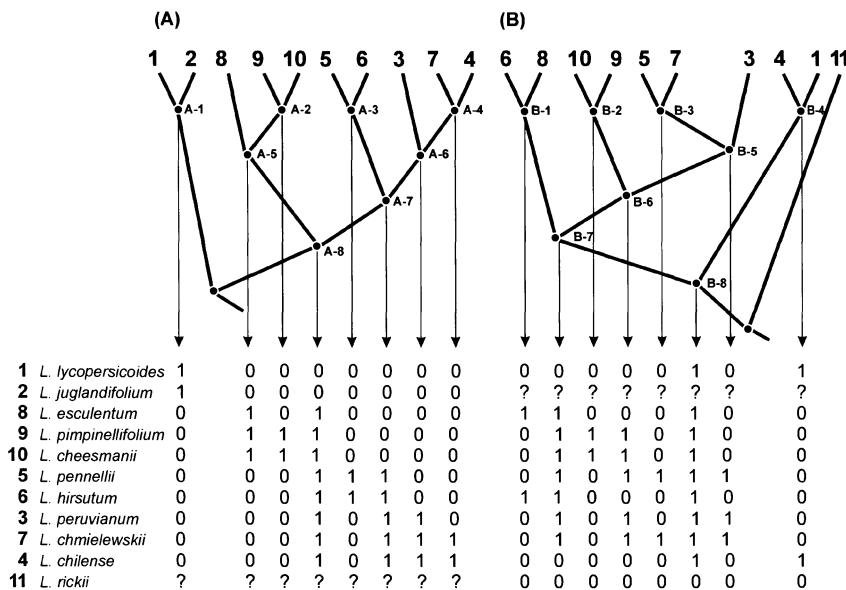


Figure 1. Example of the MRP coding procedure using two source trees of species of tomato (*Lycopersicon*). The left tree (A) represents a chloroplast DNA phylogeny. The right tree (B) represents a mitochondrial DNA tree. In tree (A), *L. rickii* is missing, whereas in tree (B) *L. juglandifolium* is missing. The resulting matrix (below) is analyzed using parsimony to yield the supertree. The source trees have been adapted from Palmer and Zamir (1982) and McLean and Hanson (1986), respectively.

symbol (e.g., ? as in the example; Figure 1). If the matrices were not rooted with the same taxon or with a dummy outgroup, stop and go back to step 2.

5. Analyze the combined matrix by parsimony, with or without “character” weighting (see below for discussion), to yield the MRP supertree. If a tree-inference method other than parsimony is used, the tree might be referred to simply as the MR supertree (dropping the “P” for parsimony) or replacing the P by the method (e.g., as in MRC, with C for compatibility).

2.1 Mathematical and algorithmic basis of the method

We appreciated from the beginning that MRP has some difficult mathematical properties (Baum, 1992). Some properties have since been clarified, whereas others await rigorous investigation still.

Like most other supertree methods, MRP requires that nodes within each source tree be arranged hierarchically. This is achieved only when trees are rooted (McMorris, 1985; Böcker *et al.*, 2000; Steel *et al.*, 2000). Source trees can be rooted on a species held in common to all trees (Baum, 1992), or on a dummy all-zero outgroup (Ragan, 1992a).

Nodes ordered hierarchically within a tree are not independent of each other. For a bifurcating tree, penultimate nodes (i.e., those subtending exactly two termini or leaves) constitute the extreme case: if such a node subtends taxa 1 and 2, then the two terminal nodes (leaves) are determined (one is taxon 1 and the other is taxon 2). This simple case does not occur in practice because, by convention, terminal nodes are not coded into the MRP matrix (Ragan, 1992a). But, at the next-higher hierarchical level, a node subtending taxa 1, 2 and 3 must give rise to a node that subtends taxa 1 and 2, or 1 and 3, or 2 and 3. The conditional probability of each, given the existence of the parent node (1, 2, 3), is 0.33. In the opposite (basal) direction, given the existence of a node that subtends two termini, say (4, 5), its immediate parent must subtend either three or four taxa, including taxa 4 and 5. Non-terminal nodes are not determined fully in either direction by their hierarchically adjacent neighbors (e.g., node A–1 by nodes A–2 to A–4 in Figure 1), but their range of allowable states is constrained variously. Indeed, these constraints propagate hierarchically, albeit in attenuated fashion, up and down the tree, and are not restricted to immediately adjacent nodes. The argument can be extended easily to trees with polytomies as well. It is worth considering whether these constraints violate the conditions under which the tree-reconstruction method (e.g., parsimony) is applied or interpreted validly.

Purvis (1995a) attributed certain biases in MRP supertree construction to this non-independence of nodes. Ronquist (1996) argued that these biases arise not from the presence of redundant information (indeed, he denies that additive binary matrices contain redundant information), but rather from the different relative sizes of source trees, and proposed tree-based weighting schemes that should compensate for the bias. Bininda-Emonds and Bryant (1998) recommend that any compensation should be node-based, not tree-based. Altogether, these and other biases might arise from the relative sizes (number of nodes) of, structures (shapes) of and possible non-independence among source trees (see Wilkinson *et al.*, 2004; for an example of the latter, see Gatesy *et al.*, 2002). Most aspects of supertree bias are amenable to computational simulation for source trees of moderate size and number.

We proposed that parsimony be used to recover the MRP tree because the method implements the exact algorithmic steps required to represent a matrix as a tree. Source trees are converted initially into the alternative (equivalent) data structure (i.e., represented in matrix format); parsimony

inverts this process, converting the matrix back to the tree data structure. Although this logic is difficult to fault, it does hide an important complication: whereas each individual source matrix was internally consistent, the combined matrix is not necessarily so unless, of course, all source trees are jointly compatible. When this is not the case, all conflicts must be resolved. The justification for using parsimony to resolve conflicts among elements of the combined matrix is, for us, the same as for using parsimony to resolve conflicts among standard discrete-state characters: efficiency and information content (Farris, 1979). We are aware of the diversity of opinion among the phylogenetic systematics community as to why parsimony is to be preferred (e.g., whether evolutionary or other biological considerations are valid justifications). We suspect that such explanations apply to matrix (topological) elements, if at all, with diminished force, but leave this debate to others (e.g., Rodrigo, 1993, 1996; Bininda-Emonds *et al.* 2002; Cotton and Page, 2004; Ross and Rodrigo, 2004). There have been suggestions that conflict among elements of the combined matrix might be resolved alternatively by compatibility approaches (Ragan, 1992b; Rodrigo, 1996; Goloboff and Pol, 2002; Ross and Rodrigo, 2004) or quartet-based methods (Bininda-Emonds *et al.*, 2002; Piaggio-Talice *et al.*, 2004), or by removing the offending taxa (Wilkinson *et al.*, 2001); these alternatives have received little scrutiny, either analytically or with real data.

MRP supertrees can take much longer to compute than supertrees based on strict supertree algorithms (Sanderson *et al.*, 1998). The latter run in polynomial time, whereas the parsimony methods used to analyze the combined MRP matrix explore tree space typically using heuristics that require exponential time. However, more-efficient optimization methods have become available recently. In particular, variants of Markov chain Monte Carlo (MCMC) methods can be many orders of magnitude faster than, for example branch-and-bound algorithms, and have been implemented for binary characters in MrBayes 3.0 (Huelsenbeck and Ronquist, 2002). Advanced search strategies such as MCMC will ensure that MRP remains computationally feasible even for data sets based on large numbers of taxa (see also Ronquist *et al.*, 2004). As with parsimony analysis, the need or interpretation of all evolutionary models required for MCMC remains an open question that we leave for others to debate.

3. Arguments advanced for and against the method

MRP was and is a unique method and does not, except in extreme or trivial cases, reduce to a variant of any other method. Rohlf (cited by Baum, 1992)

considered MRP to be “a new kind of consensus”, although Baum (1992) distinguished it from consensus approaches, and Ragan (1992a) allowed only that it resembles consensus methods by incorporating topological information from derived trees. Bininda-Emonds and Bryant (1998) reserve the term consensus for tree-based methods, whereas MRP is node-based (see below also). However, Thorley (2000) concluded that whether or not MRP should be viewed as a type of consensus method is ultimately a matter of definition.

Rodrigo (1993) raised two major criticisms against MRP. First, he objected that the method cannot resolve conflicts by reference to primary data because it is not based on combined data sets. We (Baum and Ragan, 1993) admitted that our earlier choice of words “combining data sets” was imprecise, meaning instead that MRP was a method for “making inferences from multiple data sets”. However, Rodrigo’s core criticism remains: conflicts are resolved in MRP by reference to inferred topological features, not directly from data. This is a necessary feature of indirect supertree methods (*sensu* Wilkinson *et al.*, 2001) and is part of the price to be paid for the benefits of flexibility, compactness and computability. There are many problems for which it is simply not feasible to combine evidence: one of us (Ragan) works currently with multi-genome protein-sequence data sets that, if combined into a single matrix, would have >100 rows and >1 million columns. Such data sets present difficulties owing to computational complexity (number of rows) and practicalities of data handling (number of columns), and can be approached feasibly in modular fashion only (Daubin *et al.* 2001; Roshan *et al.*, 2004).

In defense of MRP, Williams (1994), Purvis (1995a) and others have argued that MRP does in fact capture data partitions, if not data *per se*. Moreover, MRP can behave well in simulations (Bininda-Emonds and Sanderson, 2001), although it tends not to resolve polytomies as fully as does total evidence (Bininda-Emonds and Bryant, 1998).

Rodrigo’s (1993, 1996) second criticism was that MRP lacks an underlying model. We did not accept this criticism (Baum and Ragan, 1993). In a passage written by Baum, we argued that 1) gene trees ought to be regarded as character-state trees, 2) gene trees often conflict with each other and/or are incongruent with organism trees, and 3) there is thus a need to combine characters to infer organism trees (see also Doyle, 1992; Page and Charleston, 1997). Bininda-Emonds and Bryant (1998:498) challenged usage of the term “character” in this context, instead reserving the term for “attributes of organisms”. We also defended the use of parsimony as a mathematically justifiable heuristic (see above). Nonetheless, the perception has persisted that MRP and similar methods are “somewhat *ad hoc*” (Steel *et al.*, 2000; see also Novacek, 2001). Bininda-Emonds and Bryant (1998) and

Bininda-Emonds *et al.* (2002) have distinguished more recently tree-based methods (such as the various flavours of consensus) from node-based methods (such as MRP and its variants). We argue that a node-based approach to multi data set inference should look like MRP; probability-based and/or weighted extensions can be imagined. Perhaps this begs the issue, raised in this context by Bininda-Emonds and Bryant (1998), of “whether a tree is equal solely to the sum of its nodes”, but the collection of internal nodes is, at minimum, an information-rich and concise estimate.

The issue of underlying models also arises in another context. Ten years ago, phylogenetic methods were less advanced and genomes were, by default, assumed to be evolutionarily stable units the phylogenetic histories of which necessarily parallel those of the corresponding organisms. Within such a paradigm, it was reasonable to assume that single-gene or single-protein trees are, underneath the statistical noise and any bias caused by violations of our perhaps too-simplistic evolutionary models, basically congruent. Compatibility screening of candidate source trees or data sets (Bull *et al.*, 1993) seemed to be called for only in those unusual cases in which incongruence was an explicit possibility (e.g., organelle- or parasite-derived genes). Today, with sophisticated site-rate corrections available in phylogenetic software and with lateral gene transfer back firmly on the research agenda for microbial genomes at least (Doolittle, 1999; Ragan, 2001; Gogarten *et al.*, 2002), incompatibility screening might be prudent. However, its use should be determined by the biological question, and not seen as a necessarily methodological adjunct of MRP (or of any other method).

4. Modifications of MRP

Substantial scope remains for improving MRP and MRP-like supertree methods. Every step in the inference pathway might be targeted for improvement: the underlying data sets, inference of the source trees, coding of topological information, weighting of matrix elements, generation of the supertree, identifying and correcting for biases, and applying the supertree to real-life problems. Several of these steps are common to all phylogenetic methods, whereas others are specific to supertrees or MRP supertrees. We have introduced alternatives already (above) to parsimony for supertree generation; most other modifications so far have focused on coding, weighting, and allowable transformations among “character” (matrix element) states.

4.1 Coding

Motivated to remove the supposedly redundant information introduced by additive binary coding, Purvis (1995a) suggested a modification of MRP in which taxa absent from both a clade emanating from a particular node and from the sister taxon of this clade are represented by a ? instead of a 0. This modification has been dubbed “Purvis coding” (Bininda-Emonds and Sanderson, 2001). Ronquist (1996) showed that this modification reduces the information content of MRP matrices, and can lead to taxon instability in the supertree inferred from the combined MRP matrix. Bininda-Emonds and Bryant (1998) attribute this instability to removal of zeroes that, although not strictly informative in a traditional sense, nonetheless restrict the placement of taxa in the supertree. Although MRP envisions that all component trees are rooted consistently with each other, strictly speaking it is necessary only that component trees be rooted. If component trees are rooted inconsistently among themselves, then characters cannot be coded to represent ancestral (e.g., 0) and derived (e.g., 1) states.

4.2 Weighting

Ronquist (1996:253) concluded that “weight (ing) each additive binary character in proportion to the support for the corresponding grouping in the original analysis” might make the resulting supertree a better representation of the underlying data. This has been dubbed “weighted MRP” (Bininda-Emonds and Sanderson, 2001). Another approach might be to apply successive weighting, whereby characters are reweighted according to their fit to the supertree in the previous round of analysis. Baum (1992) noted that support from entire individual source trees might be assessed using Farris’s (1972, 1989) inconsistency measure of fit. Contributions from source trees might also be assessed based on *a priori* considerations (e.g., gene length). Weighting might alternatively be applied in a more-focused manner based on differential support for individual subtrees among the source trees (Bininda-Emonds and Bryant, 1998:504–505), although specific support measures might be inapplicable, or fail, in specific cases (Salamin *et al.*, 2002).

4.3 Allowable transformations among “character” states

Bininda-Emonds and Bryant (1998) considered whether an analogue of so-called Dollo parsimony might be the appropriate way to analyze MRP matrices because absence of a subtree (coded as 0) does not imply that any

feature has been lost; this approach was later termed “irreversible MRP” (Bininda-Emonds and Sanderson, 2001).

5. Merits of the method

5.1 Economy in dealing with multiple data sets, especially of molecular sequences

Because node-based information about the tree(s) derived from each data set, not the full underlying data, are retained and combined, the MRP matrix can be orders of magnitude smaller (as measured by number of columns) than the source data sets, especially for molecular sequences (Baum, 1992; Ragan, 1992a), and are correspondingly easier to deal with. Although we envisaged originally that a single best (i.e., most-likely, most-parsimonious, or shortest) tree would be selected for each source data set, equal-best trees might be represented. In extreme cases (i.e., many equal-best trees), the represented matrix might be as large as, or even larger than, the source data set (Ragan, 1992a). If n equal-best trees are represented for a single source data set, each might be weighted $1/n$ to not overwhelm contributions from other source data sets. Alternatively, one might represent only the unique nodes from the component trees (Bininda-Emonds and Bryant, 1998).

5.2 Retention of the best tree(s) for each component data set

The binary coding of each node (subtree) of each source tree preserves information about how the corresponding data set resolves relationships best among the species represented therein; each source tree (hence, indirectly, each source data set) is given a voice in the result (i.e., equal representation) without being overwhelmed by the other data sets. One of us (Baum) argues that this is akin to character trees in essence, although not in context. Conflicts are not, however, resolved by direct reference to the primary data (Rodrigo, 1996), as is possible (within a uniform data type) with TE. This was never our intention.

5.3 Ability to deal with disparate data types

Supertree methods, including MRP, remain the only method able to support inferences based on diverse data types natively, and moreover to do so in a

way in which molecular-sequence data do not (as feared as early as Kluge, 1983) overwhelm all others. In 1990, it would have required great optimism to predict that in little more than a decade, more than 100 organismal genomes would be sequenced fully, and that at least a few genes would be sequenced fully from essentially every known higher taxon. For lesser-studied organisms such as protists, it still made sense in 1990 to build composite phylogenies based on molecular sequences; discrete-state morphological, ultrastructural and biochemical data; nutrient and growth factor requirements; immunological and serological reactivities; and the like. This is much less the case today, although arguments can be made for combining nucleotide (RNA) and protein-sequence data. MRP can be useful as well for the joint analysis of fossil and extant data (Bininda-Emonds *et al.*, 2002).

5.4 Ability to deal with “missing” leaves

The past decade has witnessed little further success in development of consensus-based methods that deal informatively with partially overlapping, but incompatible source trees. Indeed, some authorities (e.g., Bininda-Emonds *et al.*, 2002) restrict usage of the term consensus to methods that combine fully overlapping source trees only. Gordon’s (1986) consensus supertree method deals informatively with partially overlapping source trees only if they are compatible (i.e., can be combined, or contained within a larger tree, without topological conflict). By contrast, MRP and some other supertree-based methods, including MINCUTSUPERTREE (Semple and Steel, 2000), require component trees to share only two leaves pairwise. Although the MRP method does not fail algorithmically until source trees are completely disjunct, the result becomes devoid of meaning below or, one suspects, at or even slightly above the limit of two.

5.5 Performance

Based on extensive simulations, MRP appears to yield good approximations to the TE approach, and its performance (in terms of accuracy) increases and often slightly surpasses that of TE when the matrix of binary factors is weighted appropriately (Bininda-Emonds and Sanderson, 2001). The effect of non-identical taxon sets is detrimental to both MRP and TE; simulation results showed that weighted MRP always outperformed TE by the criteria used in the comparison (Bininda-Emonds and Sanderson, 2001)

6. What MRP does not do

6.1 Data sets with different phylogenetic histories

MRP and other supertree methods do not sort orthologues from paralogues magically, nor reconcile phylogenetic histories that differ as a result of symbiosis, hybridization, lineage sorting, and similar phenomena (Soltis and Kuzoff, 1995; Kellogg *et al.*, 1996). Only orthologous data should be used to reconstruct gene trees, whether by supertrees or other methods. If our goal is to reconstruct species phylogeny, these gene trees can serve as component trees in supertree methods such as MRP.

6.2 Limited overlap and disjunction of species sets

The presence of overlapping but non-identical taxon sets (i.e., inconsistent sample trees; Gordon, 1986) is known as the supertree problem (Sanderson *et al.*, 1998; Steel *et al.*, 2000). Some of its aspects are formally intractable (e.g., compatibility of unrooted trees, MRP), but can be addressed in alternative ways that are, to some extent, mutually incompatible (Steel *et al.*, 2000). Put simply, the fewer cases of non-identical taxa among the source data sets, the better. The solution we provided in coding of missing taxa (above) appears satisfactory for relatively small numbers of non-identical taxa. Analytical studies and simulations are needed to assess how, and how rapidly, supertrees are weakened by increasing degrees of species non-overlap. Gordon (1986) suggested that simulations are unlikely to lead to clear-cut recommendations. Constantinescu and Sankoff (1995) have generalized Gordon's results, and Wiens and Reeder (1995) have proposed a heuristic in which the combined tree is inferred first from a complete data set, after which this tree is adjusted based on data containing the missing taxa. Simulation studies (Bininda-Emonds and Sanderson, 2001) reveal that MRP is at least as resilient as TE to increasing non-overlap of species among source trees. These studies need to be extended to discover, for example, failure modes and thresholds below which it becomes useless to elaborate a supertree by MRP or its modifications.

6.3 Equally good supertrees

We dealt above with problems that might arise when individual source data sets imply multiple equally good trees. Parsimony (and other optimization-based) approaches to analysis of the combined MRP matrix can similarly

return unresolved, or multiple equally good, supertrees (Steel *et al.*, 2000). MRP does not suggest how to deal with such cases natively. Steel *et al.* (2000; also Ross and Rodrigo, 2004) insist that all supertree methods should return only one tree, which might not be resolved fully however. Equally good supertrees from the same combined matrix would necessarily be compatible, so consensus methods would be applicable (as in Figure 4 of Baum, 1992), although at the risk of loss of resolution. Lack of resolution at the supertree level should be taken as a warning about the consistency or quality of the source data.

7. Applications of MRP in biology

MRP was first applied to molecular-sequence data by Ragan (1992a) and to combined morphological, ultrastructural and biochemical data sets for eukaryotes by Ragan (1992b). Since then, MRP has been applied (with or without modification) *inter alia* in reconstructing the phylogeny of primates (Purvis, 1995b; Purvis and Webster, 1999), carnivores (Bininda-Emonds *et al.*, 1999), placental mammals (Liu *et al.*, 2001), artiodactyls (Gatesy *et al.*, 2002; Mahon, 2004), bats (Jones *et al.*, 2002), seabirds (Kennedy and Page, 2002), *Schistosoma* (Morand and Müller-Graf, 2000), dinosaurs (Pisani *et al.*, 2002), flowering plants of order Apiales (Plunkett, 2001) and genus *Lithocarpus* (Cannon and Manos, 2001), grasses (Salamin *et al.*, 2002), legumes (Sanderson *et al.*, 1998; Wojciechowski *et al.*, 2000), and bacterial genomes (Galtier and Gouy, 1994; Daubin *et al.*, 2001, 2002). Many of these, and other, applications are discussed by Gittleman *et al.* (2004) and Moore *et al.* (2004).

8. Future directions

MRP is a framework within which source trees can be not only combined, but might also be compared with the MRP supertree and thereby with each other. One could imagine an index of agreement or disagreement, analogous to measures of character congruence in classical parsimony, according to which individual columns (nodes) or groups of nodes (e.g., those corresponding to a source tree) could be scored. This measure might represent the number of simple rearrangements required to fit a particular topological feature into the MRP supertree; or the relative cost (e.g., in number of steps) of the best tree(s) in which that node appears; or a more conventional, perhaps continuous (0 to 1) measure of congruence. Component nodes or trees that are congruent topologically with the MRP

supertree, or a subtree thereof, would be characterized by a strong fit statistic, whereas poorly fitting nodes or subtrees might be considered *prima facie* evidence for lateral transfer or other complicating processes. Because incompatible source trees can be combined into a single MRP supertree, a consistency measure of this sort might provide the first (albeit incomplete) step toward a general comparison metric for any two or more trees that share some number or proportion of species.

9. MRP software and websites

9.1 Software for computing supertrees

RadCon (Thorley and Page, 2000) supports MRP coding. r8s (Sanderson, 2003) generates combined matrix representations of trees from a tree file input. SuperTree (<http://www.tcd.ie/Botany/NS/SuperTree.html>; Salamin *et al.*, 2002;) computes MRP matrices from Nexus / Newick trees. Further supertree software is also available at <http://www.tierzucht.tum.de/Bininda-Emonds/>.

9.2 Sites for supertree information and online supertree computing

Phylogenies and trait evolution (John Gittleman):

<http://faculty.virginia.edu/gittleman/research>

Phylogenetic supertrees, links, and bibliography (Oliver Eulenstein, with Michael J. Sanderson and Daniel M. Gusfield):

<http://genome.cs.iastate.edu/supertree/>

Supertree (Rod Page):

<http://darwin.zoology.gla.ac.uk/~rpage/supertree/>

Acknowledgements

We thank Olaf Bininda-Emonds for the invitation to write this chapter.

References

- BAUM, B. R. 1990. Combining datasets for cladistic analysis. *ICSEB IV, Fourth International Congress of Systematic and Evolutionary Biology. Program, p. 13; Abstracts. Affiliated Session: NT-24.* University of Maryland, USA. Unpaginated.
- BAUM, B. R. 1992. Combining trees as a way of combining datasets for phylogenetic inference, and the desirability of combining gene trees. *Taxon* 41:3–10.
- BAUM, B. R. AND RAGAN, M. A. 1993. Reply to A. G. Rodrigo's "A comment on Baum's method for combining phylogenetic trees". *Taxon* 42:637–640.
- BININDA-EMONDS, O. R. P. AND BRYANT, H. N. 1998. Properties of matrix representation with parsimony analyses. *Systematic Biology* 47:497–508.
- BININDA-EMONDS, O. R. P., GITTLEMAN, J. L., AND PURVIS, A. 1999. Building large trees by combining phylogenetic information: a complete phylogeny of the extant Carnivora (Mammalia). *Biological Reviews* 74:143–175.
- BININDA-EMONDS, O. R. P., GITTLEMAN, J. L., AND STEEL, M. A. 2002. The (super)tree of life: procedures, problems, and prospects. *Annual Review of Ecology and Systematics* 33:265–289.
- BININDA-EMONDS, O. R. P. AND SANDERSON, M. J. 2001. Assessment of the accuracy of matrix representation with parsimony analysis supertree construction. *Systematic Biology* 50:565–579.
- BÖCKER, S., BRYANT, D., DRESS, A. W. M., AND STEEL, M. A. 2000. Algorithmic aspects of tree amalgamation. *Journal of Algorithms* 37:522–537.
- BROOKS, D. R. 1981. Hennig's parasitological method: a proposed solution. *Systematic Zoology* 30:229–249.
- BULL, J. J., HUELSENBECK, J. P., CUNNINGHAM, C. W., SWOFFORD, D. L., AND WADDELL, P. J. 1993. Partitioning and combining data in phylogenetic analysis. *Systematic Biology* 42:384–397.
- CANNON, C. H. AND MANOS, P. S. 2001. The use of morphometric shape descriptors in relation to an independent molecular phylogeny: the case of fruit type evolution in Bornean *Lithocarpus* (Fagaceae). *Systematic Biology* 50:860–880.
- CONSTANTINESCU, M. AND SANKOFF, D. 1995. An efficient algorithm for supertrees. *Journal of Classification* 12:101–112.
- COTTON, J. A. AND PAGE, R. D. M. 2004. Tangled trees from molecular markers: reconciling conflict between phylogenies to build molecular supertrees. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 107–125. Kluwer Academic, Dordrecht, the Netherlands.
- DAUBIN, V., GOUY, M., AND PERRIERE, G. 2001. Bacterial molecular phylogeny using supertree approach. *Genome Informatics* 12:155–164.
- DAUBIN, V., GOUY, M., AND PERRIERE, G. 2002. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Research* 12:1080–1090.
- DOOLITTLE, W. F. 1999. Phylogenetic classification and the universal tree. *Science* 284:2124–2128.
- DOYLE, J. J. 1992. Gene trees and species trees: molecular systematics as one-character taxonomy. *Systematic Botany* 17:144–163.
- FARRIS, J. S. 1972. Abstract of compatibility clustering. *Classification Society Bulletin* 2:35.
- FARRIS, J. S. 1979. The information content of the phylogenetic system. *Systematic Zoology* 28:483–519.

- FARRIS, J. S. 1989. The retention index and the rescaled consistency index. *Cladistics* 5:417–419.
- FARRIS, J. S., KLUGE, A. G., AND ECKHARDT, M. J. 1970. A numerical approach to phylogenetic systematics. *Systematic Zoology* 19:172–191.
- GALTIER, N. AND GOUY, M. 1994. Molecular phylogeny of Eubacteria:a new multiple tree analysis method applied to 15 sequence data sets questions the monophyly of Gram-positive bacteria. *Research in Microbiology* 145:531–541.
- GATESY, J., MATTHEE, C., DESALLE, R., AND HAYASHI, C. 2002. Resolution of a supertree / supermatrix paradox. *Systematic Biology* 51:652–664.
- GITTLEMAN, J. L., JONES, K. E., AND PRICE, S. A. 2004. Supertrees: using complete phylogenies in comparative biology. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 439–460. Kluwer Academic, Dordrecht, the Netherlands.
- GOGARTEN, J. P., DOOLITTLE, W. F., AND LAWRENCE, J. G. 2002. Prokaryotic evolution in light of gene transfer. *Molecular Biology and Evolution* 19:2226–2238.
- GOLOBOFF, P. A. AND POL, D. 2002. Semi-strict supertrees. *Cladistics* 18:514–525.
- GORDON, A. D. 1986. Consensus supertrees: the synthesis of rooted trees containing overlapping sets of labeled leaves. *Journal of Classification* 3:335–348.
- HUELSENBECK, J. P. AND RONQUIST, F. 2001. MrBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17:754–755.
- JONES, K. E., PURVIS, A., MACLARNON, A., BININDA-EMMONDS, O. R. P., AND SIMMONS, N. B. 2002. A phylogenetic supertree of the bats (Mammalia: Chiroptera). *Biological Reviews* 77:223–259.
- KELLOGG, E. A., APPELS, R., AND MASON-GAMER, R. J. 1996. When genes tell different stories: the diploid genera of Triticeae (Gramineae). *Systematic Botany* 21:321–347.
- KENNEDY, M. AND PAGE, R. D. M. 2002. Seabird supertrees: combining partial estimates of procellariiform phylogeny. *The Auk* 119:88–108.
- KLUGE, A. G. 1983. Cladistics and the classification of the great apes. In R. L. Ciochon and R. S. Corruccini (eds), *New Interpretations of Ape and Human Ancestry*, pp. 151–177. Plenum, New York, NY.
- KLUGE, A. G. 1989. A concern for evidence and a phylogenetics hypothesis of relationships among Epicrates (Boidae, Serpentes). *Systematic Zoology* 38:7–25.
- LIU, F.-G. R., MIYAMOTO, M. M., FREIRE, N. P., ONG, P. Q., TENNANT, M. R., YOUNG, T. S., AND GUGEL, K. F. 2001. Molecular and morphological supertrees for eutherian (placental) mammals. *Science* 291:1786–1789.
- MAHON, A. S. 2004. A molecular supertree of the Artiodactyla. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 411–437. Kluwer Academic, Dordrecht, the Netherlands.
- MCLEAN, P. E. AND HANSON, M. R. 1986. Mitochondrial DNA sequence divergence among *Lycopersicon* and related *Solanum* species. *Genetics* 112:649–667.
- MCMORRIS, F. R. 1985. Axioms for consensus functions on undirected phylogenetic trees. *Mathematical Biosciences* 74:17–21.
- MOORE, B. R., CHAN, K. M. A., AND DONOGHUE, M. J. 2004. Detecting diversification rate variation in supertrees. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 487–533. Kluwer Academic, Dordrecht, the Netherlands.
- MORAND, S. AND MÜLLER-GRAF, C. D. M. 2000. Muscles or testes? Comparative evidence for sexual competition among dioecious blood parasites (Schistosomatidae) of vertebrates. *Parasitology* 120:45–56.

- NOVACEK, M. J. 2001. Mammalian phylogeny: genes and supertrees. *Current Biology* 11:R573–R575.
- PAGE, R. D. M. AND CHARLESTON, M. A. 1997. From gene to organismal phylogeny: reconciled trees and the gene tree / species tree problem. *Molecular Phylogenetics and Evolution* 7:231–240.
- PALMER, J. D. AND ZAMIR, D. 1982. Chloroplast DNA evolution and phylogenetic relationships in *Lycopersicon*. *Proceedings of the National Academy of Sciences of the United States of America* 79:5006–5010.
- PIAGGIO-TALICE, R., BURLEIGH, J. G., AND EULENSTEIN, O. 2004. Quartet supertrees. In O. R. P. Bininda-Emonds (ed). *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 173–191. Kluwer Academic, Dordrecht, the Netherlands.
- PISANI, D., YATES, A. M., LANGER, M. C., AND BENTON, M. J. 2002. A genus-level supertree of the Dinosauria. *Proceedings of the Royal Society of London Series B, Biological Sciences* 269:915–921.
- PLUNKETT, G. M. 2001. Relationship of the order Apiales to subclass Asteridae: a re-evaluation of morphological characters based on insights from molecular data. *Edinburgh Journal of Botany* 58:183–200.
- PURVIS, A. 1995a. A modification to Baum and Ragan's method for combining phylogenetic trees. *Systematic Biology* 44:251–255.
- PURVIS, A. 1995b. A composite estimate of primate phylogeny. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 348:405–421.
- PURVIS, A. AND WEBSTER, A. J. 1999. Phylogenetically independent comparisons and primate phylogeny. In P. C. Lee (ed.), *Comparative Primate Socioecology*. Cambridge University Press, Cambridge.
- RAGAN, M. A. 1991. A hybrid phylogenetics based on matrix representation of trees. *Programme, Fifth Annual Meeting, Canadian Institute for Advanced Research, Program in Evolutionary Biology, Lac Delage, Québec, 10–14 August 1991*. Unpaginated.
- RAGAN, M. A. 1992a. Phylogenetic inference based on matrix representation of trees. *Molecular Phylogenetics and Evolution* 1:53–58.
- RAGAN, M. A. 1992b. Matrix representation in reconstructing phylogenetic relationships among the eukaryotes. *BioSystems* 28:47–55.
- RAGAN, M. A. 2001. Detection of lateral gene transfer among microbial genomes. *Current Opinion in Genetics and Development* 11:620–626.
- RODRIGO, A. G. 1993. A comment on Baum's method for combining phylogenetic trees. *Taxon* 42:631–636.
- RODRIGO, A. G. 1996. On combining cladograms. *Taxon* 45:267–274.
- RONQUIST, F. 1996. Matrix representation of trees, redundancy, and weighting. *Systematic Biology* 45:247–253.
- RONQUIST, F., HUELSENBECK, J. P., AND BRITTON, T. 2004. Bayesian supertrees. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 193–224. Kluwer Academic, Dordrecht, the Netherlands.
- ROSHAN, U., MORET, B. M. E., WILLIAMS, T. L., AND WARNOW, T. 2004. Performance of supertree methods on various data set decompositions. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 301–328. Kluwer Academic, Dordrecht, the Netherlands.
- ROSS, H. A. AND RODRIGO, A. G. 2004. An assessment of matrix representation with compatibility in supertree construction. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 35–63. Kluwer Academic, Dordrecht, the Netherlands.

- SALAMIN, N., HODKINSON, T. R., AND SAVOLAINEN, V. 2002. Building supertrees: an empirical assessment using the grass family (Poaceae). *Systematic Biology* 51:112–126.
- SANDERSON, M. J. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absences of a molecular clock. *Bioinformatics* 19:301–302.
- SANDERSON, M. J., PURVIS, A., AND HENZE, C. 1998. Phylogenetic supertrees: assembling the trees of life. *Trends in Ecology and Evolution* 13:105–109.
- SEMPLE, C. AND STEEL, M. 2000. A supertree method for rooted trees. *Discrete Applied Mathematics* 105:147–158.
- SOLTIS, D. E. AND KUZOFF, R. K. 1995. Discordance between nuclear and chloroplast phylogenies in the Heuchera group (Saxifragaceae). *Evolution* 49:727–742.
- STEEL, M., DRESS, A. W. M., AND BÖCKER, S. 2000. Simple but fundamental limitations on supertree and consensus tree methods. *Systematic Biology* 49:363–368.
- THORLEY, J. L. 2000. *Cladistic Information, Leaf Stability and Supertree Construction*. Ph.D. dissertation, University of Bristol, United Kingdom.
- THORLEY, J. L. AND PAGE, R. D. M. 2000. RadCon: phylogenetic tree comparison and consensus. *Bioinformatics* 16:486–487.
- WIENS, J. J. AND REEDER, T. W. 1995. Combining data sets with different numbers of taxa for phylogenetic analysis. *Systematic Biology* 44:548–558.
- WILKINSON, M., THORLEY, J. L., LITTLEWOOD, D. T. J., AND BRAY, R. A. 2001. Towards a phylogenetic supertree of Platyhelminthes? In D. T. J. Littlewood and R. A. Bray (eds), *Interrelationships of the Platyhelminthes*, pp. 292–301. Taylor and Francis, London (as cited in Bininda-Emonds *et al.*, 2002).
- WILKINSON, M., THORLEY, J. L., PISANI, D., LAPOINTE, F.J., AND MCINERNEY, J. O. 2004. Some desiderata for liberal supertrees. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic supertrees: combining information to reveal the Tree of Life*, pp. 227–246. Kluwer Academic, Dordrecht, the Netherlands.
- WILLIAMS, D. M. 1994. Combining trees and combining data. *Taxon* 43:449–453.
- WOJCIECHOWSKI, M. F., SANDERSON, M. J., STEEL, K. P., AND LISTON, A. 2000. Molecular phylogeny of the “temperate herbaceous tribes” of papilionoid legumes: a supertree approach. In P. Herendeen and A. Bruneau (eds), *Advances in Legume Systematics* 9:277–298. Royal Botanic Garden, Kew.

Chapter 2

AN ASSESSMENT OF MATRIX REPRESENTATION WITH COMPATIBILITY IN SUPERTREE CONSTRUCTION

Howard A. Ross and Allen G. Rodrigo

Abstract: Matrix representation with compatibility (MRC) identifies the largest set of mutually compatible characters (maximum clique) in combined data sets of trees represented by additive binary coding. The supertree can be determined directly from this clique, without recourse to arguments involving parsimony and homoplasy. We compared the powers of MRC and matrix representation with parsimony (MRP) to construct a supertree reliably by simulating sets of consistent and inconsistent sample trees derived from an original model tree. Under stringent definitions of success, MRC and MRP were successful with data sets having larger numbers of trees (>7 – 10), each with substantial overlap ($>50\%$ of all taxa). Overall, MRP was slightly more successful than MRC in recovering the original model tree. Identifying a maximum clique is subject to the NP-hard constraint so that fast computers and efficient software are needed for MRC to be a practical tool in the immediate future. Weakly compatible splits used in the construction of splits graphs might offer an alternative method and warrant further investigation.

Keywords: compatibility; matrix representation; maximum clique; MRC; MRP; splits graphs; supertree

1. Introduction

Although the scope of phylogenetic studies is growing rapidly, it is agreed widely that many phylogenetic relationships will only ever be investigated or revealed by the combination of phylogenetic trees to construct supertrees (Sanderson *et al.*, 1998; Soltis and Soltis, 2001; Bininda-Emonds *et al.*,

Table 1. Some algorithms appropriate for combining phylogenetic trees under different circumstances.

Consistent trees	Inconsistent trees
BUILD (Aho <i>et al.</i> , 1981)	MRP (Baum, 1992; Ragan, 1992; Baum and Ragan, 2004)
Strict Consensus (Gordon, 1986)	MRP variants:
MRP (= MRC)	coding procedures (Purvis, 1995; Wilkinson <i>et al.</i> , 2001; Semple and Steel, 2002) flip supertrees (Chen <i>et al.</i> , 2001; Burleigh <i>et al.</i> , 2004) irreversible parsimony (Bininda-Emonds and Bryant, 1998) MRC (Rodrigo, 1996)
	MINCUTSUPERTREE (Semple and Steel, 2000)
	Semi-strict supertrees (Goloboff and Pol, 2002)

2002). Biologists have two primary motivations in pursuing supertrees. In the first, and older, motivation, they wish to bring together the character-state trees indicated by individual genetic or morphological characters to achieve a better estimate of the phylogeny. More recently, as the volume and range of genetic data grows, there is the realization that constructing phylogenies of large numbers of taxa using very large data sets is impractical computationally. Our scientific appetite has outstripped our experimental capabilities. Supertrees provide a means of combining existing phylogenies to bring together taxa in novel ways and to link isolated phylogenetic trees.

Several methods of supertree construction are available and the investigator will choose one according partly to the nature of the trees that are to be combined (Table 1; see also Bininda-Emonds *et al.*, 2002).

Matrix representation with parsimony (MRP) has become a popular method, in part perhaps because it is easy to employ. MRP, which involves additive binary coding and maximum parsimony (MP), was devised by Baum (1992) and Ragan (1992) for combining phylogenetic trees. Each binary variable represents one of the partitions on the phylogenetic tree, and is therefore a single “hypothesis of relationships”. The use of parsimony as a heuristic to summarize these hypotheses has been criticized because of the lack of meaning in such a procedure (Rodrigo, 1993, 1996; Slowinski and Page, 1999; Cotton and Page, 2004). Goloboff and Pol (2002) have also criticized the method for generating groups that are unsupported by any combination of, or even contradicted by some of, the source trees. For character data, MP finds the phylogenetic tree(s) having the minimum number of evolutionary events; it might be appropriate to apply MP in this context to construct a tree that requires the minimum number of homoplasies. Homoplasies are inferences of real evolutionary events —

convergences, parallelisms or reversions. When the characters are not real and tangible but are instead hypotheses of relationships, as in MRP, then it is hard to see what homoplasies signify, and by extension, why minimizing them is an acceptable approach.

Additive binary codings of phylogenetic trees encode logical relationships relating to clade membership (Bininda-Emonds and Bryant, 1998), not directed or undirected character states. Inconsistent trees can contain phylogenetic noise: they might reflect gene trees and not species trees, combine phylogenetic estimates from traits with different or non-uniform rates of change, or suffer from sampling error. Nevertheless, we work under the assumption that all these trees contain information about a true underlying phylogeny of species, and that this true tree is what we wish to estimate. Consequently, it might be more appropriate first to identify those binary characters that are logically compatible and not contradictory, and then to find the tree that they encode. Compatibility analysis (Meacham and Estabrook, 1985) is a technique that discovers the largest set of mutually compatible characters. Two characters are compatible if and only if there exists a tree on which evolutionary changes to the characters can be proposed without invoking homoplasy. Thus matrix representation with compatibility (MRC) has the potential to identify the largest set of hypotheses of relationships that the source trees allow us to postulate, and to construct a supertree using only that set. This feature might make MRC more desirable than MRP for inferring phylogeny and constructing supertrees. It has been pointed out recently by Goloboff and Pol (2002) that MRC does not necessarily discover all uncontradicted groups implied by a set of source trees, and we will return to this point later.

Although MRP has become a popular technique (Bininda-Emonds and Sanderson, 2001), questions remain from a theoretical perspective whether it is appropriate for the reconstruction of phylogenetic relationships. MRC has received scant attention and needs to be assessed as a method for the construction of supertrees. Whereas many have assessed or disputed the validity of particular supertree methods by counterexample, we use a simulation approach, generating data sets by Monte Carlo methods and estimating empirically the abilities of both MRP and MRC to recover the true tree. Our aim is to assess the practicality of using MRP and MRC in real world cases. In particular, our approach is driven by a very simple operational criterion that we discuss below, but state now: a supertree method is successful if it finds only a single tree that is identical to the true tree.

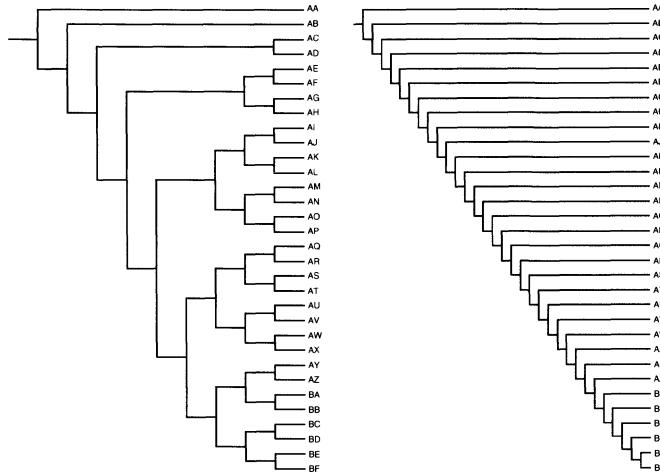


Figure 1. Examples of the original trees used in simulations. The balanced (left) and unbalanced (right) trees of 32 taxa have been rooted arbitrarily at the first taxon for presentation purposes only.

2. Methods

2.1 Generation of consistent sample trees

Consistent sample trees were generated by selecting a proportion of the taxa in a predefined (“true”) tree randomly and pruning them subsequently from the tree. Trees containing 32, 64, or 128 taxa and having either a balanced (symmetric) or unbalanced (fully asymmetric) topology (Figure 1) were used as starting trees. To generate a data set of N sample trees, the first tree was created by pruning a proportion of randomly selected taxa from the tree. For subsequent trees in the data set, taxa to be pruned were selected randomly from the set of all hitherto unpruned taxa until every taxon had been pruned once. Thereafter, taxa were selected at random without consideration for how many times they had been pruned previously. Consequently, each taxon was represented in at least one tree. One hundred replicate data sets were constructed for each combination of number of taxa, proportion deleted (pruned), and number of sample trees analyzed. In some instances indicated in the results, the number of replicate data sets was reduced when we excluded those cases where the taxa in the sample trees did not form a connected set containing all taxa. Each set of sample trees was coded using matrix representation, as described by Baum and Ragan (2004), with each

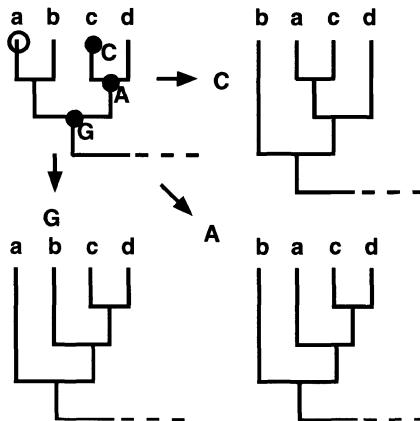


Figure 2. Generation of an inconsistent tree. A tree node is selected at random (e.g., node a; open circle, upper left) and is made the sister node to a randomly chosen cousin node (C, upper right), aunt node (A, lower right) or grandmother node (G, lower left).

sample tree being rooted arbitrarily at the first taxon before coding of the splits or other manipulations were performed.

2.2 Generation of inconsistent sample trees

Inconsistent sample trees were generated by first pruning a randomly selected proportion of the taxa in a predefined (“true”) tree, as for consistent trees, and then applying a single rearrangement on a randomly selected node. Trees containing 32 or 64 taxa, and having a balanced topology, were used as starting trees. An attempt was made to use trees of 128 taxa, but only a few combinations of parameter values were tested because of the constraints of computer processor speed and memory.

The following procedure was followed in rearranging each sample tree to make it inconsistent with the true tree (Figure 2). First, a node in the tree was chosen randomly as the candidate to be moved. A relationship was then chosen randomly, and the node having that relationship to the candidate node was identified as the destination node. For any given node (e.g., taxon a in Figure 2), three types of relationships were recognized: cousin, aunt and grandmother. Working back towards the root, the ancestors are successively the mother and grandmother (labeled G) nodes. Thence following a different branch towards the tips of the tree, the descendants of the grandmother are successively the aunt (labeled A) and cousin (labeled C) nodes. Where there were multiple potential destination nodes having the specified relationship, as with cousin nodes, one was chosen randomly. Then the candidate node

was made the sister node to the destination node. The tree manipulations were implemented in Perl using the BioPerl Tree module (Stajich *et al.*, 2002).

The rearrangement was rejected and the candidate or destination node was reselected if any of the following was true:

1. The candidate node was at the root of the tree.
2. The mother (immediately ancestral) node was at the root of the tree.
3. The aunt node was a terminal node when the cousin relationship was selected.
4. The grandmother node was at the root of the tree when the grandmother relationship was selected.

Although Figure 2 shows rearrangements involving a terminal node (taxon) as the candidate, all nodes were eligible for selection. Only a single character was altered when the transformation moved a randomly selected node to become sister taxon to either its aunt or grandmother node, whereas two characters were altered when it became sister to a cousin node. The rearrangements ranged in magnitude from shifting a terminal node to an adjacent “cherry” to reordering of the major clades near the base of the tree. They produced sample trees with inconsistencies comparable to those cited by Wilkinson *et al.* (2001) and Goloboff and Pol (2002) to challenge MRP.

2.3 Character compatibility and cliques

Characters are said to be *compatible* when they support, or are consistent with, a particular phylogenetic tree; in other words, when they can both be free of homoplasy on at least one tree. In the pairwise comparison of binary characters, we can identify compatibility operationally when no more than three of the four pairs of possible character states (0, 0), (0, 1), (1, 0) and (1, 1) occur across the taxa under consideration. It has been proven that a group of binary characters that are pairwise compatible are collectively compatible (Estabrook *et al.*, 1976). Consequently, if two binary characters are each compatible with a third, and compatible with each other, then the three are compatible (i.e., there is at least one tree where the three characters can be free of homoplasy). Sets of compatible characters are known as *cliques*.

When we combine the matrix representations of trees that do not have exactly the same membership, five additional character pairs can occur: (?, ?), (0, ?), (1, ?), (?, 0), and (?, 1). When computing the pairwise compatibility of binary characters across taxa, these character-state pairs involving one or two missing values were ignored, as in the CLINCH

software (Estabrook *et al.*, 1977; <http://www.geocities.com/RainForest/Vines/8695/clinch62.zip>).

The characters generated by the binary representation of a phylogenetic tree are necessarily consistent with that tree, and so they are collectively compatible. When the binary representations of two or more trees are considered jointly, two different situations can occur. First, when the trees are consistent, all the binary characters will be compatible. Consequently, the set of compatible characters will equal the set of characters. Second, when we combine the matrix representations of inconsistent trees, some of the binary characters in the data set will be mutually incompatible and several cliques can occur. Although the number and size of the cliques will depend on the extent to which branches of the trees are consistent, two conceptually different groups of cliques can be identified. The first group has just one member: the set of characters that is fully compatible with the true tree and hence supports it. To distinguish it from other cliques, we will term it the *best clique*. These characters individually are compatible with every character in the matrix representation of the true tree. They represent those portions of the topology of the true tree that are represented correctly in the sample trees. From the best clique, the true tree could be recovered potentially. Knowledge of the membership of the best clique, although hidden from the investigator, is the goal of a compatibility analysis. The second group comprises those cliques observable by compatibility analysis. One is the largest or *maximum clique* of compatible characters. This observable clique must be at least as large as, but need not contain, the best clique. In fact, there can be several maximum cliques, only one of which might be equivalent to the best clique. The maximum clique can also be larger than the best clique, containing all or some of its characters. When sets of trees of increasing consistency are considered, the largest or maximum clique increases commensurately to become the set of all characters when full consistency occurs.

The effect of missing entries when the matrix representations of inconsistent trees are combined can be understood by considering three three-taxon trees (Figure 3; adapted from figure 7 of Goloboff and Pol, 2002), each a sample drawn from the true and unknown tree (A, (B, (C, D))). Sample trees I and II are consistent with, but tree III is inconsistent with, the true tree. Applying the methods described above to assess the compatibility of characters in the sample trees (Figure 3, left), and ignoring character state pairs involving a missing value, leads one to conclude that all characters in the sample trees are compatible, as indicated by an asterisk in the row labeled “max”. Thus, the maximum clique comprises all characters in the data set. This maximum clique describes three trees ambiguously: each of the sample trees with the missing taxon placed at the root. By contrast, when

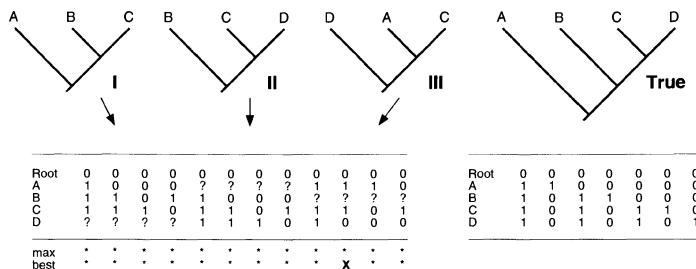


Figure 3. Missing data and compatibility. The matrix representations of three sample trees I, II, and III (left), two of which (I and II) are consistent with the true tree (right). Membership of the maximum clique (“max”) and of the best clique (“best”) is indicated by an asterisk.

each character in the sample tree data set is compared with every character in the matrix representation of the true tree (Figure 3, right), we find that the tenth character is not consistent with the true tree, as indicated by an X in the row labeled “best”. Therefore, the best clique in the sample data set comprises 11 characters, one fewer than in the maximum clique. This best clique describes the true tree unambiguously. When character-state pairs with missing values were ignored, incompatibility between the characters of trees I or II with those of tree III was obscured with a resulting overestimate of character compatibility. Thus, the maximum clique can exceed the best clique, and increasingly so when the inconsistency between sample trees is disguised by poor taxon overlap resulting in missing binary characters. The main goal of this study is to identify the circumstances under which the maximum clique provides a misleading estimate of the best clique and hence the true tree.

2.4 Criteria for success

Supertree construction is motivated usually by pragmatic considerations relating to extracting additional information about phylogenetic relationships from several trees. Consequently, we have taken a pragmatic approach when assessing methods of supertree construction. Most investigators (e.g., Bininda-Emonds and Sanderson, 2001) have assessed the accuracy of tree reconstruction using MRP on a continuous scale using measures of tree similarity. In part, this is because they have accepted multiple equally most parsimonious trees as valid outcomes in tree construction, from which they extract the strict consensus supertree. But what does one do with multiple supertrees? There are two possible solutions here: 1) try to narrow down the choices to a single supertree or 2) build a consensus supertree. But in doing

the latter, we move from a second-order analysis (i.e., the construction of supertrees from component trees) to a third-order analysis (i.e., the summarizing of several supertrees). With each step, interpretation becomes an increasing problem: what does it mean to have a consensus tree of supertrees? After all, if a supertree is a summary of component trees, then isn't a consensus of supertrees, a summary of summaries? In any case, how should we construct such a consensus? There is a relatively large literature on consensus trees and the differences between these methods (Wilkinson, 1994). But, of course, supertree methods themselves can be viewed as generalizations of consensus methods (Sanderson *et al.*, 1998; Steel *et al.*, 2000). Does consistency require that we apply the same algorithm to summarize the supertrees as well? Because no non-arbitrary method exists for the researcher to choose among these equivocal trees (or, for that matter, to summarize them), we consider ambiguity a failure of the method to construct the true tree. Readers might see this as an overly stringent criterion; if so, they can view our analyses as worst-case results. We, of course, believe that a single, correct supertree relieves users from the agony of choice, and consequently, is an admirable criterion for success.

Of course, in standard phylogenetic analyses, polytomies or consensus topologies are accepted usually, not as accurate reconstructions of phylogenetic relationships, but as compromises for handling conflicting signal or lack of resolution. Nonetheless, it is probably fair to say that our desire is always that our methods will deliver the one true tree. If this is a Utopian goal, it is still possible to measure the success of any method by its ability to attain this outcome.

To reiterate, the goal in constructing a supertree is to develop a tree that summarizes the sequence of cladogenetic events accurately, and, if we disregard hard polytomies, there is only one such tree. Consequently, we have used a strict criterion of success, namely finding a single tree identical to the original tree.

3. Results

3.1 Consistent trees

The matrix representations of consistent trees form a single set of mutually compatible characters. Because these characters form a single clique that is identical to the maximum clique, an assessment of MRC is therefore equivalent to an assessment of MRP. The assessment of both MRC and MRP was therefore performed by generating sample trees through the random

deletion of taxa from a standard tree and then estimating the frequency empirically with which MP recovered the original tree successfully (i.e., found a single MP tree that matched the original tree).

3.1.1 Preliminary investigations

We investigated two software programs, PHYLIP v3.57c (Felsenstein, 1989) and PAUP* (Swofford, 2002), for their suitability for assessing the application of MRP or, in this case, the equivalent MRC, to the estimation or reconstruction of the true tree from which the synthetic data sets had been derived. Our goal was to determine whether either program could recover the true tree reliably from consistent sample trees.

First, we used PHYLIP to find the largest clique of characters, in this case all characters, and the trees that they suggest. The CLIQUE module could not be used because it does not handle cases with missing data. Instead, on the advice of the user documentation, we used the MIX module with the T (threshold) option set to the value 2 (Felsenstein, 1981). To see if the estimates improved with deeper searching, the order in which the taxa were added was randomized using the J (jumble) option and rerun different numbers of times.

Second, we used PAUP* to search for the MP tree(s). The data were analyzed either using stepwise addition to create the starting tree or using the true tree as the starting tree, from which the search began. Branch swapping was performed using the TBR algorithm and, in the case when stepwise addition was used to generate the starting tree, only a single replicate was applied.

Figure 4 summarizes the results from three combinations of parameter values, all performed on balanced trees: A) 32 taxa, 25% deleted, two sample trees; B) 64 taxa, 50% deleted, three sample trees; and C) 128 taxa, 75% deleted, 10 sample trees. A single replicate was used in each case. Each estimate of the true tree was compared with the original by calculating the symmetric difference index (SDI). In case A, PAUP* discovered 125 equally most parsimonious trees, whereas MIX discovered 106 trees. These sets of trees had very similar distributions of SDI and both included a tree identical to the true tree ($SDI = 0$). In case B, the SDI distribution of trees discovered by PAUP* varied little at increasing numbers of trees saved. Similarly, the distribution of trees discovered by MIX did not change with an increased number of re-estimates. In case C, PAUP* found a set of trees having a unimodal SDI distribution, whereas MIX found a set having a multimodal distribution. One of the modes in the MIX set coincided with the mode of the PAUP* set, but the MIX set was broader and flatter. The two software programs found sets of trees having different SDI distributions, but neither

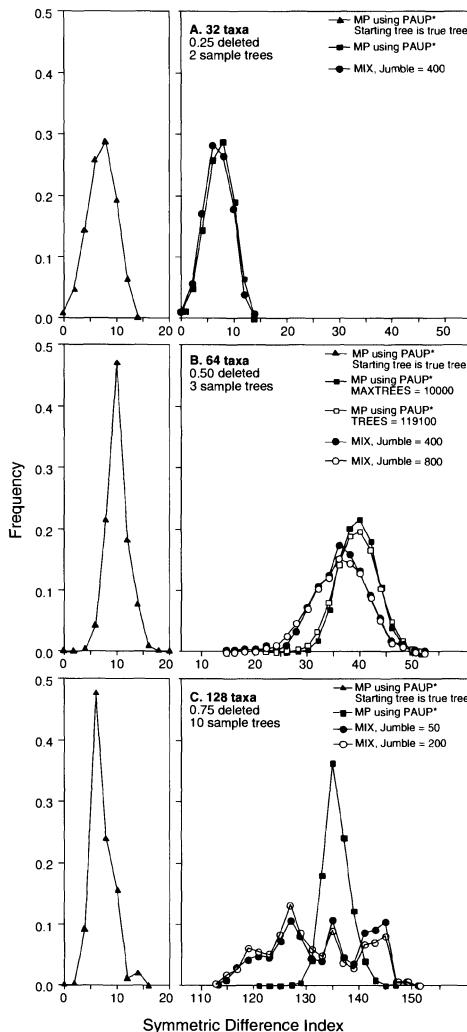


Figure 4. Frequency distributions of the symmetric difference index (SDI) between the true tree and that estimated by maximum parsimony (MP) using PAUP* or using MIX. Note that the scale for the SDI axis differs for each case.

set included the true tree. However, when the MP analysis was performed in PAUP* with the true tree as the starting tree, it discovered a set of equally most parsimonious trees that included the true tree in each case. In case A, the set of 125 equally most parsimonious trees was identical to that found when the starting tree was obtained by stepwise addition. However, the sets of 162 400 and 63 700 equally most parsimonious trees found in cases B and

C, respectively, had narrower SDI distributions than, and differed from, the sets obtained when pairwise addition was used to find the starting tree.

These results indicated that the true tree was recoverable from the combined matrix representations of sample trees of the types described above. However, the data sets contained substantial ambiguity sufficient to suggest large numbers of equally most parsimonious trees. Except in case A, both PAUP* and MIX were unable to recover the true tree without constraint. This points to the dangers inherent in heuristic branch swapping and is perhaps a salutary lesson: in all practical instances, the user has no knowledge about the true tree, and must rely on the results of the possibly flawed and incomplete results of an analysis that offers no guarantee that it has found all trees, or even the best tree.

3.1.2 Finding the true tree

Our goal in this analysis was to estimate the probability of recovering the true tree from a matrix representation of several consistent sample trees, all derived from that true tree. Our criterion for success for both MRC and MRP was finding only one MP tree (using PAUP*) when the true tree was used as the starting tree. If two or more trees were found, then one would not have an unambiguous estimate of the true tree, as found in the preliminary investigations described above. Our results comprise the percentage of replicates, under each combination of parameter values, when success occurred.

The results for balanced and unbalanced trees, and for 32-, 64- and 128-taxon trees are nearly identical (Figures 5 and 6). A very high rate of success occurred in the region characterized by small proportion deleted and large number of sample trees. The minimum number of sample trees for which (near) 100% success occurred increased from five with a proportion deleted of 0.125, to 30 with a proportion deleted of 0.50, beyond which full success was not observed. Very low success occurred in the region of large proportion deleted and small number of sample trees. The minimum number of sample trees for which any effectively non-zero success occurred rose from three with a proportion deleted of 0.25, to 20 with a proportion deleted of 0.625, and no success was observed at greater proportions deleted. Between these two regions of high and low success is a relatively steep transition. This “cliff” has a curved shape; at lower proportions deleted, increasing the number of sample trees resulted in a substantial increase in the rate of success. However, the payback diminished rapidly beyond a proportion deleted of 0.375, and full success cannot be expected beyond 0.625 for the types of data sets considered here.

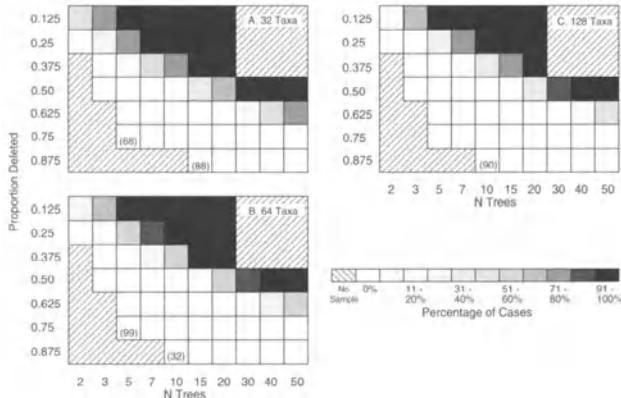


Figure 5. Estimate of the rate of success of recovering the true balanced tree from the combined matrix representations of several consistent sample trees all derived from true trees with 32 (a), 64 (b), and 128 (c) taxa. N Trees is the number of consistent sample trees in the dataset and Proportion Deleted is the fraction of the true tree that was removed in the creation of each sample tree. For each combination of N Trees and Proportion Deleted, 100 replicate datasets were analyzed, except as indicated by (n).

3.2 Inconsistent trees

In this section, we compare the effectiveness of MRC in recovering the true tree from a sample of inconsistent trees to that of MRP. We constructed data sets comprising the combined matrix representations of several smaller trees, each generated by pruning selected taxa randomly from the underlying tree and then subjecting them to a single topological rearrangement to render them incompatible with the true tree (see Section 2.2). This simulates the vagaries of taxon sampling and non-uniform rates of evolution, which contribute to uncertainty in phylogenetic trees.

In assessing MRC, we ask whether the maximum clique provides a reliable estimate of the true underlying tree by comparing it with the best clique, which comprises those characters that are fully compatible with the true tree. In a real situation, we do not know which characters comprise the best clique; we can only obtain an estimate of it in the maximum clique. In a simulation study, however, we can make a direct comparison between the best clique and the maximum clique. It is also desirable to identify just one clique as an estimate of the true tree. Multiple maximum cliques are less desirable because we lack a non-arbitrary method for choosing among them. We will judge the true tree to be recoverable to the extent that its topology is recorded in the best clique when a single maximum clique is found that contains exactly the same characters as the best clique.

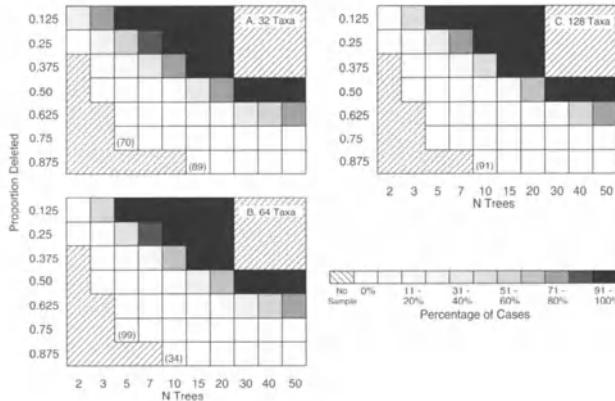


Figure 6. Estimate of the rate of success of recovering the true unbalanced tree from the combined matrix representations of several consistent sample trees all derived from true trees with 32 (a), 64 (b), and 128 (c) taxa. N Trees is the number of consistent sample trees in the dataset and Proportion Deleted is the fraction of the true tree that was removed in the creation of each sample tree. For each combination of N Trees and Proportion Deleted, 100 replicate datasets were analyzed, except as indicated by (n).

The maximum clique problem (Pardalos and Xue, 1994) is known to be NP-hard (Garey and Johnson, 1979). Consequently, operational issues become significant in applications involving maximum cliques such as this. Therefore, we also assess the usability of this method in a real-world computing environment.

3.2.1 Finding the true tree — MRC

Cliques were identified in each data set using a modified version of the CLINCH v6.2 software, developed by George Estabrook and Kent Fiala (Estabrook *et al.*, 1977; <http://www.geocities.com/RainForest/Vines/8695/clinch62.zip>). This software takes a branch-and-bound approach to the iterative search for cliques. The algorithm, which was developed empirically but has not been published (K. Fiala, pers. comm.), forms the basis of the CLIQUE module in PHYLIP together with the algorithm in Bron and Kerbosch (1973). To implement CLINCH, we converted the software from FORTRAN to Perl. The resulting Perl software was validated by comparing its output with that from a compiled version of the original FORTRAN code. The test data included both the example data that accompanies the CLINCH distribution and other synthetic data sets with or without missing values. In informal tests, the Perl CLINCH software performed very well in both speed and thoroughness of search compared to a more recent algorithm (Östergård, 2002), although we cannot discount the possibility that the difference in

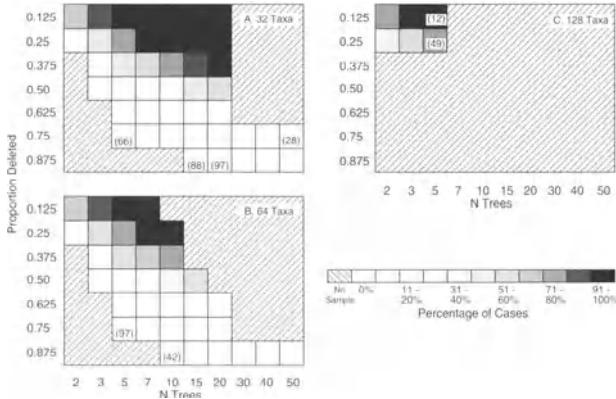


Figure 7. The frequency of cases where the size of the maximum clique was the same as the size of the best clique, comprising characters fully compatible with the original tree, for true trees with 32 (a), 64 (b), and 128 (c) taxa. N Trees is the number of inconsistent sample trees in the dataset and Proportion Deleted is the fraction of the true tree that was removed in the creation of each sample tree. For each combination of N Trees and Proportion Deleted, 100 replicate datasets were analyzed, except as indicated by (n).

performance is due entirely to the quality of the Perl code used to implement the new algorithm.

The size and number of maximum cliques varied considerably across combinations of parameter values. When compared with the best clique, the size of the maximum clique varied in a consistent manner (Figure 7). In the region characterized by small proportion deleted and large number of sample trees, the two types of cliques were the same size. At larger proportions deleted or fewer sample trees, the maximum clique was larger than the best clique more frequently. Equal clique sizes were observed only rarely at proportions deleted greater than 0.50. The results for data sets based on trees with 32, 64 and 128 taxa are very similar within the reduced sampling regime imposed by computational limitations.

The number of maximum cliques found varied from one to thousands depending on the combination of parameter values. A single maximum clique occurred frequently in two situations: 1) in association with small proportion deleted and large number of sample trees, and 2) in association with large proportion deleted and small number of sample trees (Figure 8). There is a region, running diagonally from small proportion deleted and number of sample trees to large values of both these parameters, where a single maximum clique was uncommon or did not occur. The patterns for trees with 32, 64 and 128 taxa show very similar trends, again within the limitations of our sampling regime.

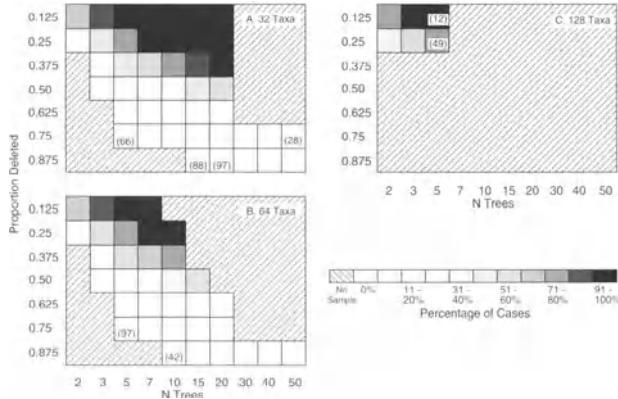


Figure 8. The frequency of cases where only one maximum clique was found, for datasets derived from true trees with 32 (a), 64 (b) and 128 (c) taxa. N Trees is the number of inconsistent sample trees in the dataset and Proportion Deleted is the fraction of the true tree that was removed in the creation of each sample tree. Sample sizes are the same as in Figure 7.

The co-occurrence of these two desirable attributes, a maximum clique having the same size as the best clique and a single maximum clique, results in the situation where the maximum clique is identical to, and recovers, the best clique. This situation was limited largely to the region characterized by small proportion deleted and large number of sample trees (Figure 9). Near zero occurrence of this situation was observed at greater proportions deleted or fewer sample trees. Separating these two areas is a region of sharp decline in the frequency of occurrence. Near 100% occurrence was observed above a minimum of seven sample trees at 0.125 proportion deleted and above 15 sample trees at 0.25 proportion deleted. The best clique was almost never recovered at proportions deleted of 0.50 and larger. Again, within the limits of our sampling regime, concordant results were obtained for all three tree sizes.

The probability of recovering the true tree from a set of inconsistent sample trees is the product of two separate probabilities. The first is the probability that the maximum clique is identical to the best clique, comprising those characters that are fully compatible with the true tree (Figure 9). The second is the probability of resolving the topology of the true tree fully from the best clique. The second probability can be inferred from our estimates of success for consistent sample trees (Figure 5) while accounting for the reduction in the number of characters through the exclusion of incompatible characters when dealing with inconsistent trees. The relative size of the best clique decreased only slightly as the proportion

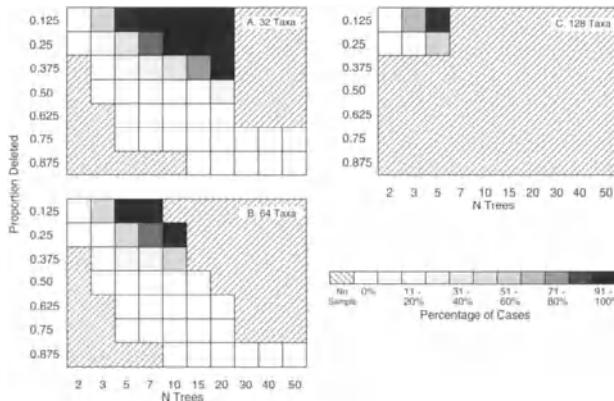


Figure 9. The frequency of cases where only one maximum clique was found, and it had the same size and membership as that for the best clique, for true trees with 32 (a), 64 (b), and 128 (c) taxa. N Trees is the number of inconsistent sample trees in the dataset and Proportion Deleted is the fraction of the true tree that was removed in the creation of each sample tree. Sample sizes are the same as in Figure 7.

deleted increased (Table 2). At the lowest proportion deleted (0.125), 98% of all binary characters were compatible with the true tree. At the highest proportion deleted (0.875), a full 91% of compatible characters remained in the data sets. Because the reduction in the data sets arising from the exclusion of incompatible characters is much less than the differences in size between adjacent data sets in Table 2, the second probability, that of recovering the true tree from a data set of consistent characters using MRP, is estimated adequately by the empirical probabilities shown in Figure 5. However, the distribution of these probabilities, especially the region of near 100% success, overlaps the region of high success in Figure 9 so that the distribution of the product is nearly identical to that in Figure 9. The transition between success and failure is sharpened, but the combinations of parameter values for which there is a high expectation of success remain the same as in Figure 9.

3.2.2 Time to Find the Maximum Clique

The time taken to find the maximum clique varied significantly across the different combinations of parameter values (Figure 10). Although the analyses were performed on different models and configurations of desktop computers that were used opportunistically as they became available, the data in Figure 10 have been selected to reflect the predominant configuration used to analyze data sets created for either 32- or 64-taxon trees.

Table 2. The size of the best clique (characters fully compatible with the true tree) as a proportion of all binary characters in the dataset. Each dataset comprises a set of inconsistent sample trees each derived from the original balanced tree of 32 taxa. D = proportion of taxa deleted from each tree.

D	<i>N</i> sample trees in data set									
	2	3	5	7	10	15	20	30	40	50
0.125	0.98	0.98	0.98	0.98	0.98	0.98	0.98			
0.250	0.97	0.97	0.97	0.97	0.97	0.97	0.97			
0.375		0.97	0.97	0.97	0.97	0.97	0.97			
0.500		0.96	0.96	0.96	0.96	0.96	0.96			
0.625			0.95	0.95	0.95	0.95	0.95			
0.750				0.93	0.93	0.93	0.93	0.93	0.93	0.93
0.875						0.91	0.91	0.91	0.91	0.91

A consistent pattern of computational constraint involving processor speed and memory was experienced. On computers configured with 128 MB RAM, all available installed memory was exhausted for data sets with approximately 1000 binary characters, and the system switched to paging virtual memory. This transition slowed the computation rate so dramatically that single replicates were not completed within ten days of elapsed time. These cases were abandoned and their timing results are not included in the following discussion. This threshold or ceiling is expected to depend on the efficiency of the software, on memory management, and on the installed memory. Improvements are expected to be linear or polynomial with improvements in the computing environment.

Where the installed memory was not exhausted, the mean time required to find the maximum clique rose exponentially with increasing number of characters (Figure 10a, b). The mean times required to find the maximum clique for 64-taxon trees were greater than for 32-taxon trees, perhaps reflecting the difference in processor speed of the computers used. The logarithm of the standard deviation of the time to find the maximum clique rose linearly with the logarithm of the mean, indicating that the variability in the computation time also increased exponentially with the number of characters (Figure 10c, d). Variation in computation time was approximately constant at small values of the mean, perhaps because elapsed time was measured to the nearest second, a scale too coarse to estimate variability accurately at short elapsed times.

These results indicate that as the number of characters increases by the addition of more or larger sample trees, both the mean and the variability in the time to find the maximum clique increase exponentially.

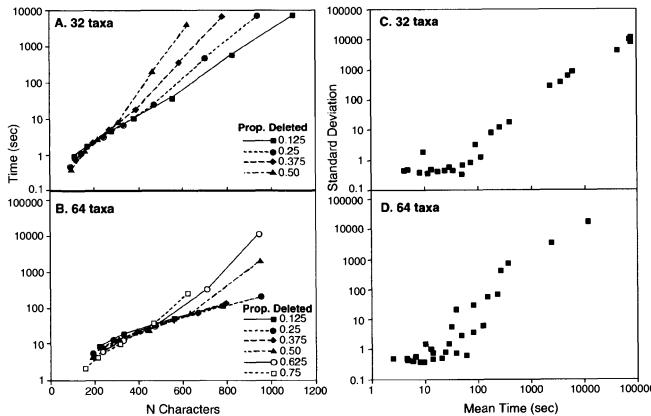


Figure 10. The time required to find the maximum clique. The left-hand graphs show the mean time as a function of the number of binary characters for 32- (a) or 64-taxon (b) trees. Prop. Deleted is the proportion of the taxa in the true tree that was deleted in generating the sample trees. The number of characters increases with the number of sample trees included in each data set. The right-hand graphs show the standard deviation of the time to find the maximum clique as a function of the sample mean for 32- (c) and 64-taxon trees (d). Estimates for 32-taxon trees (a, c) were obtained on a desktop computer with a 1.8 GHz Pentium i686 processor and 265 MB RAM. Estimates for 64-taxon trees (b, d) were obtained on desktop computers with 730 MHz Pentium i686 processor and 128 MB RAM.

3.2.3 Finding the true tree — MRP

MRP reconstructed the true tree successfully, as with previous analyses, in the region characterized by small proportion deleted and large number of sample trees (Figure 11). Results were nearly identical for all three sizes of trees investigated. Near zero success was observed at greater proportions deleted or fewer sample trees. Near 100% success was observed above a minimum of seven sample trees at 0.125 proportion deleted and above a minimum of 40 sample trees at 0.50 proportion deleted. As before, there was a sharp transition between the regions of success and failure.

3.2.4 Comparison of MRC and MRP

The power of MRC to recover the true tree was compared with that of MRP by the ratio of the respective rates of success (Table 3). For any combination of parameter values, the probability of success using MRC is the product of the probabilities given in Figures 5 and 9. Both techniques delivered near

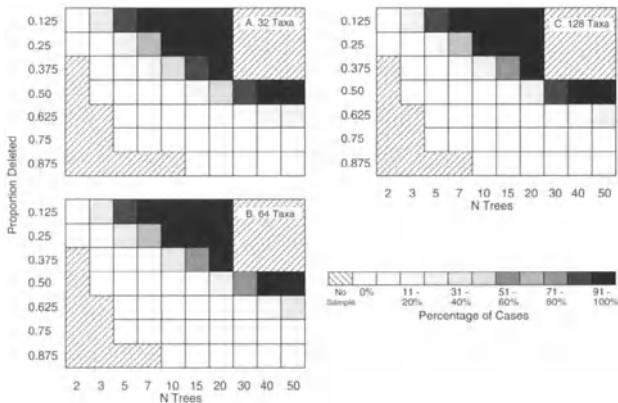


Figure 11. The frequency of cases where the original true tree was reconstructed by MRP from datasets of inconsistent sample trees derived from a balanced tree with 32 (a), 64 (b), and 128 (c) taxa. The results are based on the same datasets used in Figure 7. N Trees is the number of inconsistent sample trees in the dataset and Proportion Deleted is the fraction of the true tree that was removed in the creation of each sample tree. For each combination of N Trees and Proportion Deleted, 100 replicate datasets were analyzed.

Table 3. The power of MRC relative to MRP in reconstructing the original balanced tree from inconsistent sample trees. Comparisons were made where the success rate for MRP exceeded 50% (bold) or 10% (italic), and where an estimate of success was obtained for both methods. N taxa = number of taxa in true tree; D = proportion of taxa deleted from each tree.

N taxa	D	N sample trees in data set								
		2	3	5	7	10	15	20	30	50
32	0.125		0.74	0.95	0.98	1.00	1.00	1.00		
	0.250				0.72	0.88	0.88	0.98	0.99	
	0.375					0.59	0.52	0.65	0.90	
	0.500							0.13	0.26	
64	0.125		0.81	1.00	0.98					
	0.250				0.72	0.81	0.88			
	0.375						0.54			
128	0.125		1.00	0.84						
	0.250			0.89						

100% success when there were many sample trees each with a high degree of overlap (Table 3, bold numbers). When the success rate began to decline (Table 3, italic numbers) at lower numbers of trees or reduced overlap, MRC failed more rapidly than did MRP. Overall, MRP outperformed MRC in the probability of recovering the true tree from inconsistent source trees. The

range of parameter values over which a comparison could be made was limited by computational constraints and problems of tractability.

4. Splits graphs

One can suggest reasonably that if there are disagreements among the source trees, then building a single supertree could be inappropriate. Instead, one might wish to highlight these inconsistencies. Split decomposition (Bandelt and Dress, 1992a, b) attempts to display such inconsistencies in the data using a graph. More precisely, with binary-coded characters, where each character represents a partition or split of the taxa, the method of d -splits based on Hamming distances selects sets of weakly compatible characters to construct the graph. The method has been implemented in SplitsTree v2.4 (Huson, 1998), which we used in the following simulation. We chose for analysis three data sets having combinations of parameter values for which there was no, mixed or full success in recovering the true tree with MRP and MRC. The splits graph based on the Hamming distances in the first data set (i.e., no prior success) resolved none of the structure of the true tree, whereas the splits graph for the third data set (i.e., full prior success) resolved all of the clades in the true tree fully (results not shown). Figure 12 compares the results obtained for a data set with parameter values for which MRP and MRC had approximately 80% success in recovering the true tree. In the particular case chosen, neither MRP nor MRC were successful, and both of the reconstructed trees have the clade (AU, (AV, (AW, AX))) rather than the clade ((AU, AV), (AW, AX)) found in the true tree. The splits graph (Figure 12c) is very similar in topology to the MRP and MRC trees (Figure 12b), except that it infers a polytomy, (AU, AV, (AW, AX)), in this clade. The relationships supported by weakly compatible splits, as indicated by double parallel lines, reconstruct the remainder of the true tree accurately.

5. Discussion

Although there was a strong effect of the number of source (= sample) trees and of the tree overlap ($\approx 1 - \text{proportion deleted}$) on the probability of successful recovery the original tree using MRC, the interaction of tree number and overlap dominated the results. With consistent trees (Figures 5 and 6), a sharp transition separated a region of nearly constant success, characterized by many source trees each with substantial overlap, from a region with persistent failure to recover a single, true tree, characterized by

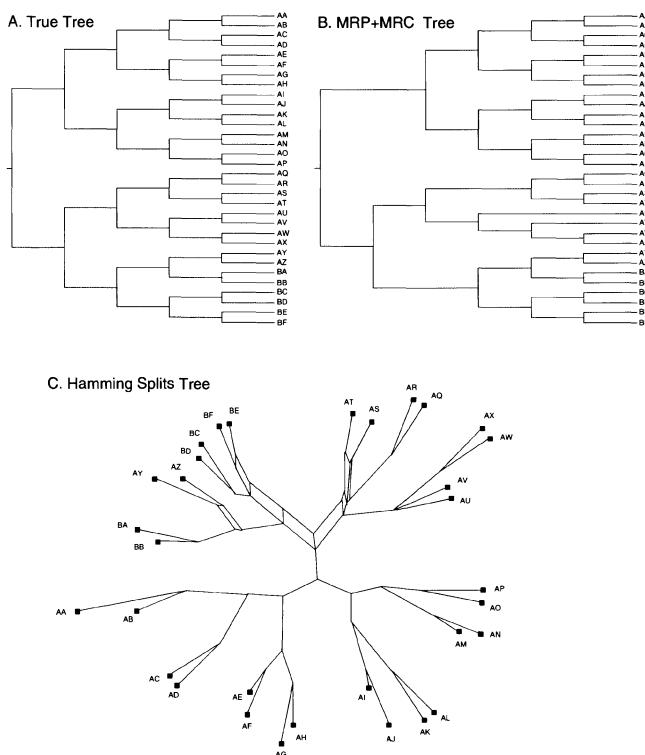


Figure 12. Comparison of the true tree (a), the tree reconstructed by MRC and MRP (b), and the splits graph of Hamming distances (c) for a dataset of inconsistent trees based on 32 taxa, with 0.125 deletion proportion and five sample trees.

fewer source trees each having less overlap. A similar pattern was found for inconsistent trees (Figure 9), but the transition was shifted to a higher number of trees and greater overlap between trees. This strong interaction between tree number and overlap was also observed when MRP was applied to the data sets of inconsistent trees (Figure 11), as found previously (Bininda-Emonds and Sanderson, 2001). In the cases of both consistent trees and inconsistent trees analyzed with MRP, a proportion deleted of 0.50 was the upper threshold for 100% rate of success. This threshold was lower for inconsistent trees analyzed by MRC at 0.25, but we could not test data sets at the combination of parameter values at which the higher threshold was reached for reasons of tractability. The minimum number of source trees at which 100% success occurred ranged from five for consistent trees to seven for inconsistent trees analyzed with MRC or MRP, and at the lowest proportion deleted (0.125) in all cases.

The topology of the trees (symmetric versus asymmetric) had little effect on the recovery of the true tree from consistent source trees (Figures 5 and 6), contrary to predictions of some authors (e.g., Purvis, 1995). Wilkinson *et al.* (2001), working with inconsistent source trees, found evidence of bias in MRP towards relationships in more asymmetric source trees. Nevertheless, on the basis of our results with consistent source trees, we decided to ignore the influence of tree topology when we came to investigate inconsistent source trees in favour of putting more resources into examining differences based on tree size.

The success of MRC in recovering a single true tree was independent of the size of the original tree. With consistent trees, there was no discernible difference in the success rate among 32-, 64- and 128-taxon trees (Figures 5 and 6). For inconsistent trees (Figure 9), the distributions of success were congruent, although the scope of the comparisons was reduced at larger tree sizes. This contrasts with a prior finding that a four-fold increase in the size of inconsistent source trees decreased the accuracy of the resulting MRP supertree significantly (Bininda-Emonds and Sanderson, 2001). As the size of inconsistent source trees increases, so too does the disagreement or noise contained in their binary coding. MRC, by identifying and excluding such noise in the form of incompatible characters, might render the estimated supertree independent of the size of the source trees. However, this does not make MRC immune to the problem of size bias (reviewed by Bininda-Emonds *et al.*, 2002). A binary data set in which a particular source tree is overrepresented will, when analyzed using MRP, give greater weight to the larger source tree in a manner analogous to majority rule. In MRC, the first step is to identify the mutually compatible characters. Because the characters derived from any single tree are necessarily compatible, a source tree of disproportionate size will generate characters that can bias the composition of the maximum clique. Such a bias would be lessened if the large and small source trees were complementary rather than contradictory in structure.

MRC should also be subject to biases arising from the non-independence of data sets, as when two or more source trees are derived either directly or indirectly from the same character data sets (reviewed by Bininda-Emonds *et al.*, 2002). Species relationships inferred by these underlying data sets will be duplicated in the source trees, and these estimates of species relationships will receive undue weighting if the source trees are combined.

A reliable estimate of the true tree can come only by discovering a unique maximum clique with membership equal to the best clique. But with MRC, we often have a situation analogous to finding multiple equally most parsimonious trees with MRP. We observed two very different patterns of variation in the size and number of maximum cliques in the data sets derived from inconsistent sample trees. In all data sets, the size of the maximum

clique is equal minimally to that of the best clique. A larger maximum clique indicates that its characters support a tree different from the true tree. A supertree derived from such a clique would be inconsistent with the true tree. A maximum clique equal in size to the best clique represents a tree having the same number of evolutionary events as the true tree, but it might have a different topology. The incidence of the maximum clique being equal in size to the best clique was related to a strong interaction between tree number and overlap, with high frequency associated with many sample trees and high overlap between trees (Figure 7). There can also be multiple maximum cliques within a data set, all or all but one of which support trees other than the true tree. A single maximum clique then is desirable, for it indicates unequivocal support for a single tree. However, a single maximum clique arose in two different situations: when large numbers of source trees with a high degree of overlap were used, and, conversely, when low overlap occurred among fewer source trees (Figure 8). The second set of conditions was associated with maximum cliques larger than the best clique. Thus, relying solely on discovering a unique maximum clique can lead to constructing the wrong tree.

When the same clade occurs in multiple source trees, the associated binary character will occur an equal number of times in the combined matrix representation. These multiple characters provide an implicit weighting for clades based on their frequency in the source trees. For example, if nine sample trees contain (A, (B, C)) and one contains (B, (A, C)), there will be nine characters supporting the former and only one supporting the latter order of branching, thereby influencing the composition of the maximum clique.

Our results indicate that the discovery of a single maximum clique will be associated with the accurate reconstruction of the true tree in a particular region of parameter-space only. Elsewhere, MRC constructs a tree having an incompletely resolved or wrong topology. For example, Goloboff and Pol (2002; their figure 2) reported that the maximum clique obtained from two inconsistent source trees generated a tree containing groups that were as often supported as contradicted in the source trees. Given that the two source trees represented a total of 16 taxa and had one-eighth and three-eighths of the full taxon set missing, respectively, our results (Figure 9) indicate that successful recovery of the true tree was not to be expected. All trees arising in this circumstance would be partly unresolved at best and misleading at worst.

In assessing the accuracy of MRP, Bininda-Emonds and Sanderson (2001) evolved DNA sequences along predefined tree topologies, generating increasing incongruence among source trees by increasing the rate of evolution. We did not set out explicitly to investigate variation in

incongruence. We made source trees inconsistent by shifting a randomly selected node to a different nearby location in the sample tree. The effect of the transformation on the topology ranged from shifting a terminal “cherry” to an adjacent cluster to repositioning large, basal clades. The transformations changed either one (aunt or grandmother target nodes) or two characters (cousin node) (Figure 2). However, because characters representing more basal clades encode the presence of more taxa than do characters representing terminal nodes, alteration of these characters might have a greater effect on the composition of the maximum clique. We attempted to simulate the average effect of inconsistency by randomization and replication. Average inconsistency among source trees, as measured by the proportion of all characters in a data set that comprise the best clique (Table 2), remained nearly constant for each proportion deleted over the range of sample trees. However, this measure fell with increasing proportion deleted, perhaps because the changed character(s) represented a greater proportion of the characters in trees with fewer taxa. Consequently, changes in the power of MRP and MRC owing to reduced overlap of trees might be confounded by increasing incongruence among source trees.

The powers of MRC and MRP to recover the true tree from data sets of inconsistent trees were very similar (Figures 9 and 11, respectively). However, when the success rate of MRC was expressed as a proportion of the success rate for MRP, we see that MRP is more successful over a wider range of parameter values, especially in the transition from near 100% success to near 100% failure (Table 3). Our criterion for success, the recovery of only the true tree, distinguished those situations where the supertree was resolved fully and reliably. Consequently, inconsistencies, which led to enlarged or multiple equally most parsimonious trees, also generated enlarged or multiple maximum cliques. Our criterion appears to have presented equivalent challenges to MRP and MRC, with MRP being the slightly more powerful method, all other things being equal.

The use of ultracliques has been proposed recently by Goloboff and Pol (2002) as an alternative to MRP in constructing supertrees. This technique finds the set of characters where each possible subset is compatible with each possible subset from the entire matrix. The result is a semi-strict consensus tree containing all the groups, and only those groups, that are implied by some combination of the source trees and contradicted by none. Goloboff and Pol provided a heuristic for finding the ultraclique, which they showed to be reliable in that it did not generate spurious groups and recovered all groups implied by the source trees. However, the technique is not guaranteed to generate a fully resolved tree, and the results of their tests bear strong similarity to ours for both MRP and MRC. The average percentage of nodes missing from the true tree in supertrees reconstructed by

ultracliques was less than 1% only at lower proportions deleted (0.25 and 0.33) and at larger numbers of source trees (≥ 5 and 8 respectively). The information content of combined tree matrices declines necessarily along the many trees-high overlap / few trees-low overlap diagonal. The method of ultracliques, like MRP and MRC, shows declining power to reconstruct the full tree as you move along this information axis. Whereas MRP and MRC will generate potentially erroneous or misleading reconstructions from weakly informative data sets, ultracliques might offer a more reliable estimate of the congruent phylogenetic content of the data.

Although there is no known polynomial-time solution to the maximum clique problem (Pardalos and Xue, 1994), it might still be feasible to use MRC in the analysis of real world data sets. The tractability of the problem depends not only on how the time taken to find a solution rises with the size of the data set, but also on the time required to solve “normal” problems. To assess the tractability, we can ask how long it would take to analyze bigger data sets that are well within the range of studies performed currently or contemplated for the near future (Bininda-Emonds *et al.*, 2002). Using the empirical formula for one of the timing profiles in Figure 10, we see that doubling the size of the data set increases the time required to find the maximum clique by four orders of magnitude. The implementation of the clique-finding algorithm used in this study was not optimized for speed, and other implementations or algorithms might emerge that will make the analysis of larger data sets tractable. Improvements in performance of several orders of magnitude will be needed to meet the growing demands imposed by ever more ambitious studies. Despite such potential improvements, compatibility analyses have only a limited usefulness in the long term because the time to a solution rises exponentially with the size of the problem. The ultraclique technique might offer a usable alternative for large data sets, given that the heuristic developed by Goloboff and Pol (2002) operates in polynomial time.

The use of MRC for constructing supertrees is arguably preferable to MRP on the grounds that MRC identifies the consistent and uncontradicted core of the data set and excludes those nodes that cannot coexist logically with others. In this sense, it represents a consensus, identifying the underlying structure on which the source trees agree. In an operational context however, it was slightly less powerful than MRP at recovering the true tree unambiguously. One of us has criticized MRP previously on the grounds of interpretability (Rodrigo, 1993, 1996; see Section 1). We believe that the ends should not justify the means, and that the shortcomings of MRP outweigh its ability to recover the true tree. Nonetheless, we accept that other users might choose ends over means, and, if this is the case, then our results indicate that MRP is to be preferred over MRC.

The technique of splits decomposition was developed to identify phylogenetic relationships that were supported by some, but not necessarily all, characters (Bandelt and Dress, 1992a, b). When applying this method to three data sets for which we had low, mixed or high prior success with MRP and MRC in recovering the true tree, we obtained splits graphs that had similar cladistic structure to the trees constructed by other means. We conclude from this that splits graphs might be a method that will prove useful in supertree construction.

The growth in the size of supertrees is accelerating. We need reliable supertree methods that will keep pace. MRC, although desirable conceptually, is not scalable given current heuristics. Can other compatibility-based methods, such as ultracliques and splits graphs, handle larger data sets? A thorough assessment of these methods is needed.

Acknowledgements

We thank Mike Steel and Charles Semple for helpful suggestions and for providing guidance on previous work in the field. This prevented us from reinventing the wheel on several occasions. Comments from Rod Page, Pablo Goloboff, and Olaf Bininda-Emonds improved the paper.

References

- AHO, A. V., SAGIV, Y., SZYMANSKI, T. G., AND ULLMAN, J. D. 1981. Inferring a tree from the lowest common ancestors with an application to the optimization of relational expressions. *SIAM Journal of Computing* 10:405–421.
- BANDELT, H.-J. AND DRESS, A. W. M. 1992a. A canonical decomposition theory for metrics on a finite set. *Advances in Mathematics* 92:47–105.
- BANDELT, H.-J. AND DRESS, A. W. M. 1992b. Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Molecular Phylogenetics and Evolution* 1:242–252.
- BAUM, B. R. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon* 41:3–10.
- BAUM, B. R. AND RAGAN, M. A. 2004. The MRP method. In O. R. P. Bininda-Emonds (ed). *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 17–34. Kluwer Academic, Dordrecht, the Netherlands.
- BININDA-EMONDS, O. R. P. AND BRYANT, H. N. 1998. Properties of matrix representation with parsimony analyses. *Systematic Biology* 47:497–508.
- BININDA-EMONDS, O. R. P., GITTLEMAN, J. L., AND STEEL, M. A. 2002. The (super) tree of life: procedures, problems, and prospects. *Annual Review of Ecology and Systematics* 33:265–289.
- BININDA-EMONDS, O. R. P. AND SANDERSON, M. J. 2001. Assessment of the accuracy of matrix representation with parsimony analysis supertree construction. *Systematic Biology* 50:565–579.

- BRON, C. AND KERBOSCH, J. 1973. Algorithm 457: finding all cliques of an undirected graph. *Communications of the ACM* 16:575–577.
- BURLEIGH, J. G., EULENSTEIN, O., FERNÁNDEZ-BACA, D., AND SANDERSON, M. J. 2004. MRF supertrees. In O. R. P. Bininda-Emonds (ed). *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 65–85. Kluwer Academic, Dordrecht, the Netherlands.
- CHEN, D., EULENSTEIN, O., FERNÁNDEZ-BACA, D., AND SANDERSON, M. J. 2001. *Supertrees by Flipping*. Technical Report TR02-01, Department of Computer Science, Iowa State University, 226 Atanasoff Hall, Ames, IA 50011–1040, USA.
- COTTON, J. A. AND PAGE, R. D. M. 2004. Tangled trees from molecular markers: reconciling conflict between phylogenies to build molecular supertrees. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 107–125. Kluwer Academic, Dordrecht, the Netherlands.
- ESTABROOK, G. F., JOHNSON, C. S., JR., AND MCMORRIS, F. R. 1976. An algebraic analysis of cladistic characters. *Discrete Mathematics* 16:141–147.
- ESTABROOK, G. F., STRAUCH, J. G., JR., AND FIALA, K. L. 1977. An application of compatibility analysis to the Blackiths' data on orthopteroid insects. *Systematic Zoology* 26:269–276.
- FELSENSTEIN, J. 1981. A likelihood approach to character weighting and what it tells us about parsimony and compatibility. *Biological Journal of the Linnean Society* 16:183–196.
- FELSENSTEIN, J. 1989. PHYLIP – Phylogeny Inference Package (Version 3.2). *Cladistics* 5:164–166. (<http://evolution.genetics.washington.edu/phylip.html>)
- GAREY, M. R. AND JOHNSON, D. S. 1979. *Computers and Intractability: A Guide to the Theory of NP-completeness*. Freeman, New York.
- GOLOBOFF, P. A. AND POL, D. 2002. Semi-strict supertrees. *Cladistics* 18:514–525.
- GORDON, A. D. 1986. Consensus supertrees: the synthesis of rooted trees containing overlapping sets of labeled leaves. *Journal of Classification* 3:335–348.
- HUSON, D. H. 1998. SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* 14:68–73.
- MEACHAM, C. A. AND ESTABROOK, G. F. 1985. Compatibility methods in systematics. *Annual Review of Ecology and Systematics* 16:431–446.
- ÖSTERGÅRD, P. R. J. 2002. A fast algorithm for the maximum clique problem. *Discrete Applied Mathematics* 120:197–207.
- PARDALOS, P. M. AND XUE, J. 1994. The maximum clique problem. *Journal of Global Optimization* 4:301–328.
- PURVIS, A. 1995. A modification to Baum and Ragan's method for combining phylogenetic trees. *Systematic Biology* 44:251–255.
- RAGAN, M. A. 1992. Phylogenetic inference based on matrix representations of trees. *Molecular Phylogenetics and Evolution* 1:53–58.
- RODRIGO, A. G. 1993. A comment on Baum's method for combining phylogenetic trees. *Taxon* 42:631–636.
- RODRIGO, A. G. 1996. On combining cladograms. *Taxon* 45:267–274.
- SANDERSON, M. J., PURVIS, A., AND HENZE, C. 1998. Phylogenetic supertrees: assembling the trees of life. *Trends in Ecology and Evolution* 13:105–109.
- SEMPLE, C. AND STEEL, M. 2000. A supertree method for rooted trees. *Discrete Applied Mathematics* 105:147–158.
- SEMPLE, C. AND STEEL, M. 2002. Tree reconstruction from multi-state characters. *Advances in Applied Mathematics* 28:169–184.

- SLOWINSKI, J. B. AND PAGE, R. D. M. 1999. How should species phylogenies be inferred from sequence data? *Systematic Biology* 48:814–825.
- SOLTIS, P. AND SOLTIS, D. E. 2001. Molecular systematics: assembling and using the Tree of Life. *Taxon* 50:663–677.
- STAJICH, J. E., BLOCK, D., BOULEZ, K., BRENNER, S. E., CHERVITZ, S. A., DAGDIGIAN, C., FUELLEN, G., GILBERT, J. G. R., KORF, I., LAPP, H., LEHVÄSLAIHO, H., MATSALLA, C., MUNGALL, C. J., OSBORNE, B. I., POCOCK, M. R., SCHATTNER, P., SENGER, C. J., STEIN, L. D., STUPKA, E., WILKINSON, M. D., AND BIRNEY, E. 2002. The BioPerl toolkit: Perl modules for the life sciences. *Genome Research* 12:1611–1618.
- STEEL, M., DRESS, A. W. M., AND BÖCKER, S. 2000. Simple but fundamental limitations on supertree and consensus tree methods. *Systematic Biology* 49:363–368.
- SWOFFORD, D. L. 2002. *PAUP*: Phylogenetic Analysis Using Parsimony (* and Other Methods)*. Sinauer, Sunderland, Massachusetts.
- WILKINSON, M. 1994. Common cladistic information and its consensus representation: reduced Adams and reduced cladistic consensus trees and profiles. *Systematic Biology* 43:343–368.
- WILKINSON, M., THORLEY, J. L., LITTLEWOOD, D. T. J. AND BRAY, R. A. 2001. Towards a phylogenetic supertree of Platyhelminthes? In D. T. J. Littlewood and R. A. Bray (eds), *Interrelationships of the Platyhelminthes*, pp. 292–301. Taylor and Francis, London.

Chapter 3

MRF SUPERTREES

J. Gordon Burleigh, Oliver Eulensteiner, David Fernández-Baca, and Michael J. Sanderson

Abstract: We survey and present new results and techniques for the supertree method *matrix representation using flipping (MRF)*. The method resolves inconsistencies among the input trees by working with the matrix representations of the clusters exhibited by the input trees. All inconsistencies between the clusters in the matrix are resolved by a minimum number of *flips*, where each flip moves a taxon into or out of a cluster. The resulting clusters form an *MRF supertree*. We present an empirical study of MRF supertrees, where input trees for the study were selected out of a large tree set using a novel graph-theoretic sampling technique that maximizes the taxon support in the resulting supertrees. This study suggests, as do simulation studies, that MRF supertrees are relatively accurate when compared to matrix representation with parsimony supertrees, MINCUTSUPERTREES, and modified MinCutSupertrees.

Keywords: biclique; clustering; compatibility; empirical study; graph editing; phylogeny; simulation study; supertree

1. Introduction

All supertree methods must deal with incompatibilities among the input trees. We describe a relatively new approach that addresses this issue through a notion of *error correction*. Our method uses the same underlying representation of the input trees as the method of matrix representation with parsimony (MRP; Baum, 1992; Ragan, 1992; also Brooks, 1981). In this encoding, trees are represented by their cluster sets, with each cluster comprising the set of all taxa (labels) that descend from the same non-root

node. Taxa present in a cluster are scored 1, those absent in the cluster are scored 0, and those not sampled on that input tree are scored by a ?. One notion of *error* in such a cluster system is the presence of an incorrect label in a cluster or the absence of one that should be present. In the matrix representation of a tree, such errors correspond to *flips* from $0 \rightarrow 1$ or $1 \rightarrow 0$, as determined by the remaining entries in the cluster system. Because of these mistakes, input trees can be incompatible with each other, and thus, when their matrix representations are combined, the resulting matrix might no longer represent any phylogenetic tree perfectly. A natural optimization problem is to find the minimum number of flips that converts the matrix into one that is consistent with a phylogenetic tree; we call this the *matrix representation using flipping (MRF) problem*. The *MRF method* constructs error-corrected supertrees, called *MRF supertrees*, by solving the MRF problem for the matrix representation of a collection of rooted phylogenetic trees. We define the MRF problem formally in the next section. Before doing so, we contrast its philosophy with that of MRP and discuss its relationship with other supertree methods.

The *MRP problem* is to find one or more equally most parsimonious trees over all matrices that result from replacing ?s with 1s or 0s. Perhaps because of the widespread acceptance of both the philosophy underlying parsimony approaches and the mechanics of parsimony analysis, and because of the availability of algorithms for maximum parsimony (e.g., Swofford, 2002), MRP is the only method widely used to construct supertrees currently. The appeal to simplicity of parsimony is relatively easy to justify in the original tree-building problem, in which homoplasy represents additional assumptions of parallel evolutionary changes or reversals. Minimizing the number of steps on a (super)tree relative to a matrix representation of a collection of trees does not have such an obvious justification. Homoplasy on a tree derived via MRP represents incompatibilities between clades rather than between individual evolutionarily novel character states (e.g., as for substitutions in a DNA sequence). MRP counts the clade represented by a character as the basic item of evidence that is competed among trees and this can lead to odd behavior, such as the method preferring relationships based largely on the size of the clade involved (Purvis, 1995). In MRF, the item of evidence is reduced down to the individual taxon in a clade. In particular, all cells in the matrix representation having either a 1 or a 0 can be regarded as potential errors. This corresponds closely to errors in the membership of a given taxon in a clade. This correspondence is not exact, however, because the hierarchical nature of trees means that an error in one cell might be correlated with errors in more inclusive clades (as in all supertree methods that utilize matrix representation). This can result in MRF also preferring relationships based largely on the size of a clade involved (e.g., Purvis,

1995). Although we have not examined the presence or characteristics of bias in MRF methods, it is possible that MRF is subject to similar biases that affect MRP (e.g., Purvis, 1995; Bininda-Emonds and Bryant, 1998).

The MRF strategy of determining the minimum number of flips that are required to turn the matrix into one that corresponds to a tree with no homoplasy is different from goal of MRP of determining what tree, given the matrix, has the least homoplasy. One way to intuit MRF is to regard the original matrix representation as a forest of input trees, and the final compatible matrix as a representation of the supertree. The minimum number of flips separating these matrices is a non-symmetric measure of distance to compatibility based on error correction. The MRF problem then seeks the supertree closest to the set of input trees using this notion of distance.

MRF is related to other supertree construction methods. If the input trees are compatible, and hence no flips are necessary, the MRF tree is a supertree displaying all of these trees. MRP has the same property. For compatible trees, however, a more efficient (indeed, polynomial-time) approach was devised by Aho *et al.* (1981; also Henzinger *et al.*, 1999). Semple and Steel (2000) modified the former algorithm to handle incompatible input. Their *MINCUTSUPERTREE (MC)* algorithm simulates that of Aho *et al.*. Whenever the simulation encounters a conflict (and thus the original algorithm would be unable to proceed), the MC algorithm deletes a minimum amount of information from the input (essentially edges from a certain graph) to allow the computation to continue. Although the optimization criterion here is local, there are some connections with the flipping problem because, in a sense, the information that is deleted corresponds to a set of $1 \rightarrow 0$ flips in the matrix representation of the input. More recently, Page (2002) presented a version of the MC algorithm, the *modified MINCUTSUPERTREE (MMC)* algorithm that retains more of the uncontradicted information from the input trees than the MC algorithm and still runs in polynomial time.

The closest relative of the MRF problem is the *fractional character compatibility problem (FCC)* of Kearney *et al.* (1999). The latter problem is based on a similarity measure that, given a set of partitions of the taxon set and an unrooted phylogeny, evaluates how well each partition is represented by the tree. It can be shown that the similarity score equals the number of taxa minus the number of modifications that must be made to the partition for the latter to correspond exactly to some cluster in the tree. Each such modification corresponds naturally to a flip. The goal in FCC is to find a tree that has maximum total similarity with a given set of input partitions. Thus, FCC can be viewed as a version of the flip problem for unrooted input trees over the same taxon set. Kearney *et al.* (1999) showed that FCC is NP-complete and give an approximation algorithm for it. Although the definition

of FCC can be simply extended to a supertree problem (including input trees over different taxon sets), the approximation algorithm cannot.

In this chapter, we define the MRF problem formally. We then survey the theoretical properties of the problem, including its NP-completeness, its fixed-parameter tractability, its approximability in certain special cases, and its properties as a consensus method (see Section 3). Although these notions lend a degree of mathematical support for MRF supertrees, computational testing is essential. To this end, we review the results of experiments that we have conducted elsewhere on simulated and real data that compared MRF supertrees with MRP, MC, and MMC supertrees (Section 4). Our first set of experiments tested exact (and therefore exponential-time) solutions to MRF against the other methods (Chen *et al.*, 2003). The results suggest that MRF is at least as accurate as its rivals in reconstructing the true tree. Unfortunately, exact algorithms can only handle small numbers (<20) of taxa; heuristics become necessary for data sets of a realistic size. A second batch of experiments evaluated the performance of a heuristic for the MRF problem, which yielded results that were similar qualitatively to those obtained with exact solutions (Eulensteiner *et al.*, in press). Of course, simulations must make assumptions that might be unrealistic; thus, the ultimate test of the usefulness of a supertree method is its performance on real data. A first step in this direction is presented in Section 5, where we describe an empirical study on plant DNA sequence data that produced encouraging results. In this latter work, other issues emerge, most notably questions of input-tree sampling strategies, which we address through a novel graph-theoretic approach. A paper that elaborates on the sampling strategy is in preparation.

2. The MRF problem

We now introduce the basic definitions and notation to be used here, and define the MRF problem formally. Throughout this chapter, S denotes a finite set of taxa $\{s_1, s_2, \dots, s_n\}$. For a rooted tree T , we write $\mathcal{L}(T)$ to denote the leaf set of T . For $X \subseteq \mathcal{L}(T)$, $T(X)$ denotes the minimal subtree in T connecting X ; the root of $T(X)$ is the node closest to the root of T .

A *phylogenetic tree* (for brevity, a *phylogeny*) over S is a rooted tree T such that $\mathcal{L}(T) = S$, where every vertex other than the root has degree at least three. Phylogeny T displays T' if $\mathcal{L}(T') \subseteq \mathcal{L}(T)$ and T' can be obtained from $T(\mathcal{L}(T'))$ by contracting a sequence of internal edges. A *cluster* in T is a subset of S consisting of all leaves that descend from some particular node of T . Note that, by definition, S is a cluster; every cluster other than S is called *proper*.

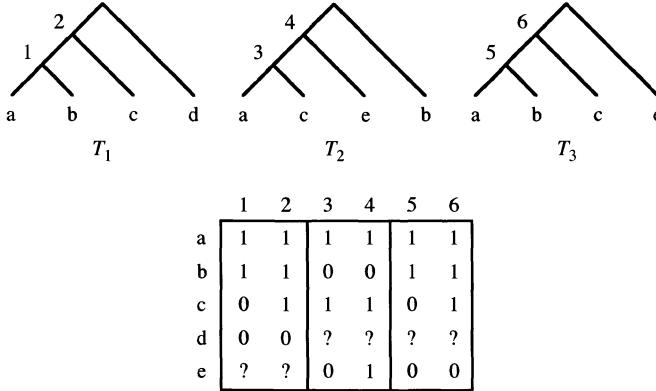


Figure 1. Top: a collection of incompatible trees $\mathcal{T} = \{T_1, T_2, T_3\}$. Bottom: $C(\mathcal{T})$; each column is labeled by the cluster to which it corresponds.

From this point forward, let $\mathcal{T} = \{T_1, \dots, T_k\}$ denote a multiset of trees, where $\cup_{i=1}^k \mathcal{L}(T_i) = S$ (that is, each T_i is a phylogeny over a subset of S). A *supertree* for \mathcal{T} is a phylogenetic tree T such that $\mathcal{L}(T) = S$. We say that \mathcal{T} is *compatible* if there is a supertree that displays each tree in \mathcal{T} . The compatibility of \mathcal{T} can be tested efficiently (Aho *et al.*, 1981; Henzinger *et al.*, 1999). However, inputs to the supertree problem are rarely compatible. The MRF and MRP methods, as well as the method of matrix representation with compatibility (MRC; Ross and Rodrigo, 2004), address incompatibility by working with a matrix representation of \mathcal{T} , which we define next.

For $p \in \{1, 2, \dots, k\}$, let $X_{p1}, X_{p2}, \dots, X_{pq_p}$ be the proper clusters of T_p in some arbitrary order. The *matrix representation* of T_p is the $n \times q_p$ matrix $C(T_p) = [c_{ij}]$, where

$$c_{ij} = \begin{cases} ? & \text{if } s_i \notin \mathcal{L}(T_p) \\ 1 & \text{if } s_i \in X_{pj} \\ 0 & \text{otherwise} \end{cases}.$$

Let $m = \sum_{i=1}^k q_i$. The *matrix representation* of \mathcal{T} is the $n \times m$ matrix $C(\mathcal{T})$ that consists of k blocks of columns, where block p is $C(T_p)$ (see Figure 1).

MRP takes the matrix representation of \mathcal{T} as a character matrix for S and seeks the phylogeny for S that requires the fewest transitions under the parsimony criteria employed (e.g., Dollo, Wagner, irreversible). By contrast, the MRF problem seeks the minimum number of changes that need to be made to $C(\mathcal{T})$ for it to correspond perfectly to some phylogeny. We now define these notions precisely.

Let $\mathcal{B} = [\beta_{ij}]$ be an $n \times m$ binary matrix and, for $j \in \{1, \dots, m\}$, let $O_j(\mathcal{B})$ denote the set of row indices i such that $\beta_{ij} = 1$. \mathcal{B} is *compatible* if there exists a phylogeny T for $S = \{s_1, s_2, \dots, s_n\}$ such that for every $j \in \{1, \dots, m\}$, there exists a cluster X in T such that $X = \{s_i : i \in O_j(\mathcal{B})\}$. The following result is well known:

Theorem 2.1 (Estabrook *et al.*, 1975; Gusfield, 1997). *A binary matrix \mathcal{B} is compatible if and only if for any pair of columns j and j' , $O_j(\mathcal{B}) \cap O_{j'}(\mathcal{B}) \in \{\emptyset, O_j(\mathcal{B}), O_{j'}(\mathcal{B})\}$.*

Let $C = [c_{ij}]$ be an $n \times m$ matrix such that $c_{ij} \in \{0, 1, ?\}$ for all i, j . A *completion* of C is an $n \times m$ binary matrix $\mathcal{B} = [\beta_{ij}]$ (that is, a matrix without question marks) such that, for all i, j , $\beta_{ij} = c_{ij}$ whenever $c_{ij} \neq ?$. C is said to be *compatible* if it has a compatible completion. Thus, a compatible completion of C exists if and only if all $?$ s in C can be changed to 0s or 1s such that for any pair of columns j and j' , $O_j(\mathcal{B}) \cap O_{j'}(\mathcal{B}) \in \{\emptyset, O_j(\mathcal{B}), O_{j'}(\mathcal{B})\}$. There are polynomial-time algorithms to test for the existence of a compatible completion and to construct one if it exists (Aho *et al.*, 1976; Pe'er *et al.*, 2000). Clearly, T is compatible if and only if $C(T)$ is compatible.

A *flip* in C is the operation of replacing an entry c_{ij} , where $c_{ij} \neq ?$, by its complement. If $c_{ij} = 0$, the flip is called a $0 \rightarrow 1$ or an *insertion flip*; if $c_{ij} = 1$, the flip is called a $1 \rightarrow 0$ or a *deletion flip*.

Figure 2 illustrates the notions of flipping, compatibility, and completion.

The *MRF problem* is, given a multiset \mathcal{T} of trees, to find the minimum number of flips needed to convert $C(\mathcal{T})$ into a compatible matrix C' . A tree T corresponding to such a C' is called the *MRF supertree*. Note that this tree is not necessarily unique. The MRF problem can be extended to the *weighted MRF problem*. Here, flips of matrix elements are weighted by numbers (e.g., to reflect differential nodal support and confidence from clustering statements). A collection of flips is weighted by the sum of their weights. The problem is to find the collection of flips with minimum weight under other collections that convert $C(\mathcal{T})$ into a compatible matrix C' .

3. Survey of theoretical results

Here we summarize some results on the computational complexity of the flipping problem as well as its consensus properties; proofs are given in Chen *et al.* (2002a, b). As before, $\mathcal{T} = \{T_1, \dots, T_k\}$ is a multiset of phylogenies such that $\bigcup_{i=1}^k \mathcal{L}(T_i) = S$.

	$C(T)$							C'					
a	1	1	1	1	1	1	a	1	1	1	1	1	1
b	1	1	<u>0</u>	<u>0</u>	1	1	b	1	1	1	1	1	1
c	0	1	1	1	0	1	c	0	1	1	1	0	1
d	0	0	?	?	?	?	d	0	0	?	?	?	?
e	?	?	0	1	0	0	e	?	?	0	1	0	0

	\mathcal{B}					
a	1	1	1	1	1	1
b	1	1	0	0	1	1
c	0	1	1	1	0	1
d	0	0	0	0	0	0
e	0	0	0	1	0	0

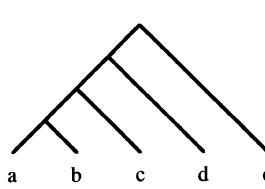


Figure 2. $C(T)$ is the matrix for the incompatible set of trees from Figure 1. C' is the compatible matrix that results from flipping the underlined entries of $C(T)$. \mathcal{B} is a completion of C' ; the tree corresponding to \mathcal{B} is also shown. Note that the placement of taxon d as ancestral to taxon e is not supported by the input trees of Figure 1; it is simply an artifact of the matrix completion chosen in this example, which is not the only one that would have led to compatibility. Note that the completion of C' can be interpreted as $? \rightarrow 1$ or $? \rightarrow 0$ flips with zero cost. Thus, there is no order between flipping and completion.

3.1 Complexity and algorithms

The *decision version* of the MRF problem is the following: given a multiset T of phylogenies and an integer k , do k or fewer flips suffice to make $C(T)$ compatible? We have the following result:

Theorem 3.1. *The decision version of the MRF problem is NP-complete, even if 1) every flip must be a $1 \rightarrow 0$ flip or 2) every flip must be a $0 \rightarrow 1$ flip. Indeed, the problem and both of its restricted versions are NP-complete even when all input trees are over the same leaf set.*

This result and others rely on the graph-theoretic formulation of the MRF decision problem, which we now review. Suppose the columns of $C(T)$ are indexed from 1 through m (recall that each column corresponds to some cluster of some tree in T). Let $G(T)$ be the bipartite graph the vertex set of which is $X \cup Y$, where $X = \{x_1, \dots, x_m\}$ and $Y = S = \{s_1, \dots, s_n\}$, and the edge set of which consists of all $\{s_i, x_j\}$ such that $c_{ij} = 1$. A Σ -subgraph in $G(T)$ is a simple path of length four in $G(T)$, the degree-one vertices of which are in

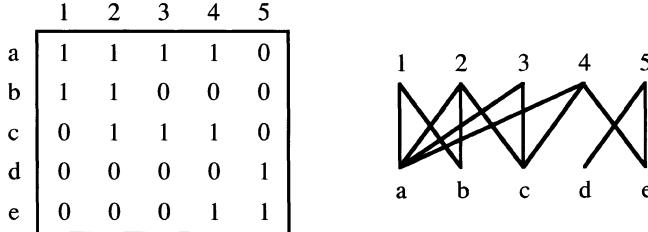


Figure 3. Left: a tuple of incompatible complete characters. Right: the corresponding character graph; note that the latter contains the induced Σ -subgraph defined by the path $\langle b, 2, c, 4, e \rangle$.

\mathcal{Y} . $G(\mathcal{T})$ is Σ -free if it contains no induced Σ -subgraph. Figure 3 depicts the bipartite graph for a tuple of incompatible complete characters.

Theorem 2.1 can be restated in graph-theoretic terms as follows:

Theorem 3.2. Suppose \mathcal{T} is a multiset of phylogenies over the same taxon set S . Then, \mathcal{T} is compatible if and only if $G(\mathcal{T})$ is Σ -free.

An *edit step* on $G(\mathcal{T})$ consists of inserting or deleting an edge. An edge deletion in $G(\mathcal{T})$ corresponds to a $1 \rightarrow 0$ flip in $C(\mathcal{T})$. If all trees in \mathcal{T} have the same leaf set, an insertion corresponds to a $0 \rightarrow 1$ flip. The MRF problem for a multiset of phylogenies over the same taxon set is thus equivalent to finding the smallest number of edit steps needed to make $G(\mathcal{T})$ Σ -free. The NP-completeness proof for the insertion-restricted MRF decision problem is by reduction from the *chain graph completion problem*, studied and proved to be NP-complete by Yannakakis (1981).

On the positive side, the MRF problem can be solved approximately within certain guaranteed bounds. A minimization problem P is *approximable within a factor of α* , for some $\alpha \geq 1$, if there exists a polynomial-time algorithm A for P such that, on any input, the cost c of the solution returned by A is within a factor of α of the cost c^* of the optimum solution; that is, $c / c^* \leq \alpha$. The following theorem relies on some general results on edge modification problems by Natanzon *et al.* (2001).

Theorem 3.3. *The MRF problem is approximable within a factor of $2d$, where d is the maximum degree of a node in $G(\mathcal{T})$ and all trees in \mathcal{T} are phylogenies over the same set of taxa. The same result holds for the version where only $1 \rightarrow 0$ flips are allowed.*

The graph-theoretic interpretation of characters is also useful for obtaining an algorithm for the version of the MRF-decision problem where a maximum number k of flips is fixed: given a multiset of trees \mathcal{T} , are at most

k flips necessary to make $C(\mathcal{T})$ compatible? The next result might be useful when k is small; it implies that the MRF-decision problem is *fixed-parameter tractable* in the sense of Downey and Fellows (1997).

Theorem 3.4. *The MRF-decision problem, as well as the versions where only $0 \rightarrow 1$ or $1 \rightarrow 0$ flips are allowed, can be solved in $O(6^k(m + n)^5)$ time when the input trees are phylogenies over the same set of taxa S .*

1.2 Consensus properties of flipping

Supertree methods are similar in spirit to consensus-tree methods, which combine information from a collection of trees that all have the same leaf set (Swofford, 1991). Thus, in principle, all supertree methods can be used to produce consensus trees. To better understand the properties of supertree methods, it is natural to investigate whether they exhibit any of the familiar properties of consensus methods. Here we present some steps in that direction. We first give some definitions.

Let X and Y be subsets of S , and T be a phylogeny for a subset of S . We say that X nests in Y , denoted $X <_T Y$, if $X, Y \subseteq L(T)$ and the lowest common ancestor of X in T is a proper descendant of the lowest common ancestor of Y in T .

Suppose X_1 and X_2 are clusters in two (possibly different) phylogenies over S . We say that X_1 and X_2 are *compatible* if there exists a phylogeny over S that has X_1 and X_2 as clusters.

Let T be a supertree for \mathcal{T} . The following definitions are adapted from Adams (1972), Bremer (1990), Böcker *et al.* (2000), and Semple and Steel (2000).

- T displays the *nestings* in \mathcal{T} if for every pair X, Y of subsets of S , $X <_T Y$ if $X <_{T_i} Y$ for every $i \in \{1, \dots, k\}$.
- T displays the *majority consensus* of \mathcal{T} if it has the following property: for every $X \subseteq S$ such that X is a cluster in over half of the trees in \mathcal{T} , then X is a cluster in T .
- T displays the *strict consensus* of \mathcal{T} if it has the following property: for every $X \subseteq S$ such that X is a cluster in every tree in \mathcal{T} , then X is a cluster in T .
- T displays the *semi-strict consensus* of \mathcal{T} if it has the following property: for every $X \subseteq S$ such that X is a cluster in some tree in \mathcal{T} and X is compatible with every cluster in every tree of \mathcal{T} , then X is a cluster in T .

It can be shown by an example that an MRF or MRP supertree T for \mathcal{T} might not display the majority consensus of \mathcal{T} or the nestings in \mathcal{T} . The latter fact implies that MRF and MRP supertrees might not display the *Adams consensus tree* (Adams, 1972) of \mathcal{T} because this consensus tree must display all nestings in \mathcal{T} (Chen *et al.*, 2003; Diao *et al.*, 2003).

On the positive side, it can be shown that all MRF or MRP supertrees display the semi-strict consensus of the input trees and, hence, also the strict consensus. Thus, whenever the input trees are compatible, T displays each of the input trees (Bryant, 2003; Chen *et al.*, 2003; Diao *et al.*, 2003). The relationship between MRF supertrees and those obtained by the method of matrix representation with compatibility (MRC; Ross and Rodrigo, 2004) is not well understood. Intuitively, however, MRF might produce more resolved trees than MRC because the latter discards incompatible characters, whereas MRF attempts to modify them so as to include them.

4. Supertree simulation study

The relative performance of supertree methods can be assessed by examining the quality of the resulting supertrees constructed from a common collection of input trees. We have conducted two types of simulation analysis elsewhere to compare the performance of the MRF method to other approaches. For input trees where the total number of taxa is small, we carried out an *exact study*, where optimal (i.e., exact) supertrees were built (Chen *et al.*, 2003). For the case where the number of taxa is large, we carried out a *heuristic study* (Eulenstein *et al.*, in press), where possibly suboptimal supertrees were constructed by heuristics. Both studies indicated that MRF supertrees are at least as accurate as, and often more accurate than, MRP, MC and MMC supertree methods. We review the results of these simulations in this section.

4.1 Experimental framework for simulation studies

The exact and heuristic studies used similar experimental designs (Chen *et al.*, 2003; Eulenstein *et al.*, in press). Model trees were generated according to a Yule birth process using the default parameters of the YULE_C procedure from r8s (Sanderson, 2003). Input trees were constructed from the model trees through three steps:

1. a gap-free alignment of DNA sequences was simulated based on a model tree,

2. the alignment was partitioned into equally sized *blocks* of consecutive columns, and
3. most parsimonious trees were constructed from each block.

The number of input trees was varied by increasing the number of blocks from two to twenty in increments of two. The length of the blocks was varied through *fixed sequence length* and *proportional sequence length* experiments. In fixed sequence length experiments, the block length is the fixed sequence length divided by a given number of input trees. Proportional sequence length experiments also started from a fixed block length, but the sequence length was instead the fixed block length times the number of input trees. To change the amount of taxon overlap between input trees, 25%, 50%, or 75% of their taxa were deleted at random. One hundred simulation replicates were performed for each combination of number of input trees, fixed and proportional sequence lengths, and deletion frequency. Finally, the accuracy of the resulting supertrees was determined by comparing the supertree to the model tree. The similarity of each supertree S and its corresponding model tree M was determined by the MAST similarity and the triplet similarity (Page, 2002) measures. The *MAST similarity* from S to M is the number of leaves of the maximum agreement subtree (MAST; Farach *et al.*, 1995) of S and M divided by the number of leaves in S . The *triplet-fit similarity* from S to M equals $1 - (d + r) / (d + r + s)$, where s is the number of identical triplets in S and M , d is the number of triplets resolved differently in S and M , and r is the number of triplets resolved in M but not in S .

4.2 Simulation results

Exact MRF and MRP supertrees can be computed only for relatively few (<20) taxa because the underlying computational problems are intrinsically hard. Simulation results demonstrate that exact MRF supertrees are almost always as accurate as exact MRP or MC supertrees (Chen *et al.*, 2003). Although the exact MRF method consistently had higher average MAST similarity scores than either exact MRP or MC supertrees, the difference was rarely significant statistically (Chen *et al.*, 2003). Still, several general trends were apparent from the data. The exact MRF and MRP supertrees were nearly always more accurate than MC supertrees (Chen *et al.*, 2003). Exact MRF supertrees generally had a higher MAST score than exact MRP supertrees, but the difference was greatest as the deletion probability increased and the taxon overlap of input trees decreased (Chen *et al.*, 2003). Although all supertree methods generally performed better with more data,

the trends in the fixed and proportional sequence length experiments were similar (Chen *et al.*, 2003).

Eulensteiner *et al.* (in press) developed an MRF heuristic that allows the construction of large, but not necessarily optimal, *heuristic MRF supertrees*. Similar to the parsimony heuristic used in PAUP* (Swofford, 2002), an initial tree is grown that is optimized further through a series of branch-swapping operations. Both processes are guided by an objective that selects one element from a “small” set of trees determined by the heuristic. Whereas a most parsimonious tree is selected in the parsimony heuristic, a tree with the minimum *flip distance* is selected in the MRF heuristic. Briefly, the flip distance from a character matrix C to a tree T is the minimum number of flips needed to convert C into a matrix representation $C(T)$. Simulations of 48- and 96-taxon trees demonstrated that heuristic MRF supertrees are at least as accurate as *heuristic MRP supertrees* (as computed using the parsimony heuristic of PAUP*) and exact MMC or MC supertrees (Eulensteiner *et al.*, in press). Heuristic MRF and MRP supertrees were more accurate than either the exact MC or MMC algorithms in all simulations (Eulensteiner *et al.*, in press). The average MAST and triplet similarity fit scores of the heuristic MRF and MRP supertrees were relatively close in all simulations, but heuristic MRF supertrees appeared generally to have slightly better scores than heuristic MRP supertrees in most simulations, especially as the taxon deletion probability increased (Eulensteiner *et al.*, in press).

4.3 Discussion of simulation studies

The results from both the exact and the heuristic study indicate that MRF supertrees are at least as accurate as MRP, MC, or MMC supertrees under a wide range of conditions (Chen *et al.*, 2003; Eulensteiner *et al.*, in press). The MC and MMC algorithms generally did not perform as well as the MRF or MRP methods (Chen *et al.*, 2003; Eulensteiner *et al.*, in press). The MRF method performed well using either exact or heuristic algorithms, indicating that it might be effective for constructing supertrees with small or large numbers of taxa (Eulensteiner *et al.*, in press). Furthermore, the MRF method appeared to be most successful relative to other supertree methods when the deletion probability was increased (Eulensteiner *et al.*, in press). The MRF method should be most distinct from the MRP method when errors exist in the input trees. If the input trees are completely without error, then the MRF and MRP methods should produce the same supertree (Chen *et al.*, 2003). More errors in input-tree construction exist when the taxon sampling is lowest and thus average branch lengths of the tree are longest. Differences in the performance of MRF and MRP supertrees were more evident in both

simulations as the deletion probability increased (Chen *et al.*, 2003; Eulenstein *et al.*, in press). The simulation results suggest that the MRF method is better able to produce an accurate supertree than the other methods under conditions in which more error can occur in the input trees.

The MAST similarity score differences between the MRF and MRP method appeared to be at least as great in the 20-taxon simulations as the 48- and 96-taxon simulations (Chen *et al.*, 2003; Eulenstein *et al.*, in press). The larger difference in MAST similarity scores even in small trees can be explained by the difference in the simulations. The 20-taxon simulations constructed trees from much smaller input-tree data sets (ranging from 50 to 500 base pairs) than the 48- and 96-taxon data sets (ranging from 1000 to 10 000 base pairs). Also, the 20-taxon simulations used a model of evolution that incorporated rate variation among sites (Chen *et al.*, 2003), which can add additional error to parsimony analyses (Sullivan and Swofford, 2001). Therefore, the differences in the experimental design of the simulations might account for additional error in the 20-taxon input trees.

5. Empirical comparison of flip-supertree methods

Although data from simulation experiments suggest that the MRF method might be useful for constructing larger supertrees, simulations make assumptions that can oversimplify the patterns of evolution. For example, the simulations assumed that the distribution of species can be described with a Yule birth process, but the actual patterns of diversification can be much more variable (e.g., Magallón and Sanderson, 2001). Also, taxon sampling of the source trees in empirical studies is unlikely to be random as it was in the simulations. Thus, more convincing arguments in favor of the MRF method could be made if its accuracy relative to other supertree methods were demonstrated on real data. This task is difficult because true evolutionary relationships are rarely known with complete certainty. Still, the relationships of certain lineages appear largely unambiguous in some cases. We designed an experiment to compare the performance of the MRF method to other supertree methods in resolving relationships among angiosperms (flowering plants). To maximize the support from a collection of input trees for each taxon in the resulting supertrees, we used a novel graph-theoretic sampling strategy. The strategy samples sets of input trees that are maximal in their number of trees times the number of taxa common to the joint overlap of all the trees. Sets of input trees were sampled from a large number of green-plant gene trees, and we compared the resulting MRF, MRP, MC, and MMC supertrees with a reference tree of generally accepted angiosperm relationships.

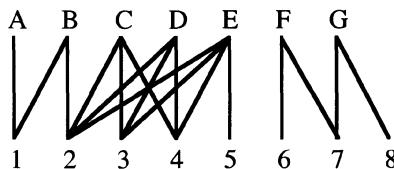


Figure 4. A bipartite graph with a maximal biclique consisting of taxon set $\{C, D, E\}$, and tree set $\{2, 3, 4\}$.

5.1 Methods for constructing plant supertrees

Under optimal circumstances, input trees would utilize as much information as possible, including the largest possible sampling of gene trees. We downloaded all available green plant sequences from GenBank as of June, 2002. To group the sequences into orthologous sequences, we first performed a pairwise BLAST search of all sequences to find groups of orthologous loci. We identified 650 distinct groups of potential orthologues with at least four taxa; these groups contained 3058 total taxa. *Arabidopsis thaliana* was the only taxon found in more than 50% of the groups, and no taxon appeared in more than 75% of the groups. We aligned the amino acid sequences within each of the groups of orthologues and found parsimony tree(s) for each alignment using protein parsimony with TBR branch swapping in PAUP* (Swofford, 2002). If there was more than one equally most parsimonious tree, we used their strict consensus as the input tree. Our sampling treats each gene tree as a potential data point for the supertree (e.g., Doyle, 1992), which avoids the problem of including trees constructed from overlapping gene data sets (e.g., Gatesy *et al.*, 2002; see also Gatesy and Springer, 2004). As discussed, simulation experiments showed that increased taxon overlap in the input trees can improve the accuracy of the resulting supertrees (Chen *et al.*, 2003; Eulenstein *et al.*, in press). In the ideal case, input trees contain the same set of taxa, and each taxon is supported by every tree; the supertree problem then becomes a consensus tree problem. The plant gene trees data set has a sparse taxon distribution and contains only small sets of taxa in which several trees overlap jointly. To allow the construction of larger and still reliable supertrees, it is desirable to have clusters of input trees that have several taxa and overlap in most of their taxa. Taxon overlap of trees can be represented through a bipartite graph $G = (X, Y, E)$, where the node set X represents the trees, and the node set Y represents the taxa from the trees considered. An edge $\{x, y\} \in E$ is drawn between a tree $x \in X$ and a taxon $y \in Y$ if and only if tree x exhibits taxon y (Figure 4).

If a subset $X' \subseteq X$ of trees jointly overlaps a subset $Y'' \subseteq Y$ of the taxa, then in the bipartite graph G every tree in X' is connected by an edge $e \in E$ to every taxon in Y' . Such a complete subgraph of a bipartite graph is called a *biclique*. For example, the tree set $\{2, 3, 4\}$ and the taxon set $\{C, D, E\}$ form a biclique in the bipartite graph depicted in Figure 4.

A biclique is *maximal* if it cannot be extended to a biclique by any edge in the bipartite graph. Thus, sets of input trees that are maximal in their cardinality times the joint taxon overlap of the trees can be found using maximal bicliques. Note that trees in a biclique can have more taxa than the biclique represents. For example, in Figure 4, the maximal biclique induced by the nodes C, D, E and $2, 3, 4$ has three taxa, whereas tree E has four taxa. Finding all maximal bicliques is at least as hard computationally as the intrinsically difficult computational *maximum edge biclique problem* (Pe'er *et al.*, 2000). Nevertheless, the specific type of problem instance (the set of gene trees) allowed us to find all maximal bicliques in our tree database in reasonable time using the enumeration algorithm of Alexe *et al.* (2000).

We discarded all maximal bicliques with fewer than five taxa and ten gene trees. We chose this cutoff based on the number of maximal bicliques that we found. Next, we used four different taxon-sampling treatments to produce the input trees. Although the gene trees in each maximal biclique share at least five taxa in common, some taxa might still be present in one or few gene trees. Therefore, to maximize the taxonomic overlap, we pruned the trees by taxa in the maximal bicliques to make sets of input trees. In the 100% biclique treatment, we pruned the trees in the biclique such that all their taxa were represented in the biclique. In the 75%, 50%, and 25% prunings, we kept taxa that were present in at least the corresponding percentage of gene trees of a biclique. Thus, each level of biclique sampling represented a different amount of taxon overlap. Next, we used outgroup rooting to root all the input trees to represent the trees as a matrix. If the input trees contained a non-angiosperm taxon, the non-angiosperm taxon was used as the outgroup. If the input tree contained only angiosperm taxa, the outgroup was chosen according to the ancestral lineages determined by recent angiosperm systematic studies (e.g., Mathews and Donoghue, 1999; Qiu *et al.*, 1999). If the ancestral species could not be determined with certainty, we made an arbitrary choice that was implemented for all input trees. Although the outgroup rooting procedure could contain error, the same criteria were used for all input trees and the same set of input trees were used for all supertree construction methods. We constructed MRF, MRP, MC, and MMC supertrees from the rooted input trees as in the heuristic simulation study of Eulensteiner *et al.* (in press).

The accuracy of the supertrees was determined by comparing the triplet similarity scores of the supertrees to a reference tree of generally accepted

angiosperm relationships. Because we acknowledge the extreme difficulty of constructing a reference tree that would be accepted as entirely correct by all systematists, we attempted to construct a tree that contained clades that are well-supported in several major studies of angiosperm relationships. With this criterion, the reference tree contained many unresolved clades (Bremer *et al.*, 1998; Soltis *et al.*, 2000). Families of uncertain general position were not included in the reference tree. We assumed that each family is monophyletic, but found that this assumption had little or no effect on the results (data not shown). Although the input trees contained sequences from all plants, the reference tree only resolved angiosperm clades. The non-angiosperms formed an unresolved clade sister to the angiosperms because approximately 90% of the sequences from the input trees came from angiosperms. The original reference tree contained several thousand taxa, but before calculating each triplet similarity score, we trimmed the supertrees and the reference tree to their common taxon sets.

5.2 Results

We found 290 maximal bicliques that met the five-taxon, ten-gene requirement. Together, the 290 maximal bicliques contained 122 of the original 650 trees, and 902 of the 3058 original taxa. No single taxon was found in every maximal biclique. The trees that composed each biclique were trimmed, and the trimmed trees in each biclique were used as input trees for supertree analyses. A total of 32 taxa were contained in all input trees that contained only the pure (100%) bicliques. None of the 32 taxa were found in every biclique. The 75% input trees also contained a total of 32 taxa, and the 50% and 25% input trees contained 34 and 48 taxa, respectively. Three taxa were found in all the 75% bicliques, and eight and 14 taxa were found in all the 50% and 25% bicliques, respectively.

The results of the empirical study largely reflect the same trends found in the simulation studies (Chen *et al.*, 2003; Eulenstein *et al.*, in press). As measured by the triplet similarity scores, all supertree methods had the most accurate trees when the supertree was constructed from a pure biclique. As the taxon sampling was extended, the average triplet similarity decreased, and there were no triplet similarity scores of one (representing total agreement of all resolved triplets with the reference tree) for supertrees made from the 50% or 25% bicliques (Figure 5). In the simulations, the accuracy of the supertree methods decreased similarly as taxon overlap decreased (Chen *et al.*, 2003; Eulenstein *et al.*, in press). Also, as in the simulations, the MC algorithm appeared to produce the least accurate supertrees in the plant data set using all sets of data (Figure 5). The MMC algorithm produced more accurate trees than the MC algorithm, but it was generally not as accurate as

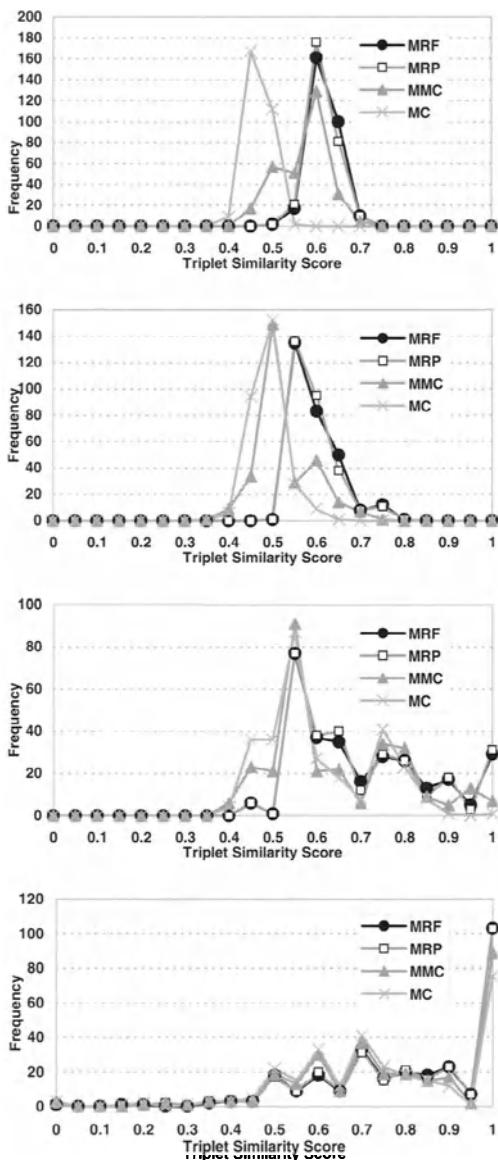


Figure 5. Frequency distributions of the triplet fit similarity of supertrees for different samplings (from top to bottom: 25%, 50%, 75%, and 100%)

the MRF and MRP heuristics in the 75%, 50%, and 25% bicliques (Figure 5). Furthermore, the performance of the MRF and MRP heuristics was largely similar at the 100%, 75%, and 50% biclique levels. However, at the

25% biclique level, MRF supertrees appeared to have a slightly higher average triple score than MRP supertrees (Figure 5).

6. Conclusion

Both exact and heuristic MRF methods appear to resolve relationships at least as accurately as other supertree methods under a wide variety of conditions. An appealing feature of the MRF method is its effectiveness at dealing with limited taxon overlap and errors. Indeed, whereas all supertree methods appear to perform better with greater taxon overlap and when combining more data, MRF supertrees perform especially well compared to the other methods when data and taxon overlap are limited. The limited taxon sampling in the plant gene trees illustrates that building supertrees that compose large parts of the Tree of Life will doubtlessly involve combining many trees with limited overlap. For example, only 36 plant taxa were contained in bicliques with at least four other taxa in at least ten trees. With the heuristic MRF method (Eulenstein *et al.*, *in press*), building MRF supertrees is now tractable for input trees containing many taxa. Thus, we propose that the heuristic MRF method is a viable alternative to the MC, MMC, and MRP methods for constructing parts of the Tree of Life.

Acknowledgements

J. Gordon Burleigh, Oliver Eulenstein, and Michael Sanderson were supported in part under NSF grant 1053164. David Fernández-Baca was supported in part under NFG grant CCR-9988348. This work was completed while J. Gordon Burleigh was at the Department of Computer Science, Iowa State University.

References

- ADAMS, E. M., III. 1972. Consensus techniques and the comparison of taxonomic trees. *Systematic Zoology* 21:390–397.
- AHO, A. V., HOPCROFT, J. E., AND ULLMAN, J. D. 1976. On finding lowest common ancestors in trees. *SIAM Journal on Computing* 1:115–132.
- AHO, A. V., SAGIV, Y., SZYMANSKI, T. G., AND ULLMAN, J. D. 1981. Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. *SIAM Journal on Computing* 10:405–421.
- ALEXE, G., ALEXE, S., FOLDES, S., HAMMER, P. L., AND SIMEONE, B. 2000. *Consensus Algorithms for the Generation of all Maximal Bicliques*. Technical Report 2000–14, Rutgers University.

- BAUM, B. R. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon* 41:3–10.
- BININDA-EMONDS, O. R. P. AND BRYANT, H. N. 1998. Properties of matrix representation with parsimony analyses. *Systematic Biology* 47:497–508.
- BÖCKER, S., BRYANT, D., DRESS, A. W. M., AND STEEL, M. A. 2000. Algorithmic aspects of tree amalgamation. *Journal of Algorithms* 37:522–537.
- BREMER, K. 1990. Combinable component consensus. *Cladistics* 9:369–372.
- BREMER K., CHASE M. W., STEVENS P. F., ANDERBERG A. A., BACKLUND A., BREMER B., BRIGGS B. G., ENDRESS P. K., FAY M. F., GOLDBLATT P., GUSTAFSSON M. H. G., HOOT S. B., JUDD W. S., KÄLLERSJÖ M., KELLOGG E. A., KRON K. A., LES D. H., MORTON C. M., NICKRENT D. L., OLSTEAD R. G., PRICE R. A., QUINN C. J., RODMAN J. E., RUDALL P. J., SAVOLAINEN V., SOLTIS D. E., SOLTIS P. S., SYTSMA K. J., AND THULIN M. 1998. An ordinal classification for the families of flowering plants. *Annals of the Missouri Botanic Garden* 85:531–553.
- BROOKS, D. R. 1981. Hennig's parasitological method: a proposed solution. *Systematic Zoology* 30:325–331.
- BRYANT, D. 2003. A classification of consensus methods for phylogenetics. In M. Janowitz, F.-J. Lapointe, F. R. McMorris, B. Mirkin, and F. S. Roberts (eds), *Bioconsensus*, pp. 163–184. American Mathematical Society, Providence, Rhode Island.
- CHEN, D., DIAO, L., EULENSTEIN, O., FERNÁNDEZ-BACA, D., AND SANDERSON, M. J. 2003. Flipping: a supertree construction method. In M. Janowitz, F.-J. Lapointe, F. R. McMorris, B. Mirkin, and F. S. Roberts (eds), *Bioconsensus*, pp. 135–160. American Mathematical Society, Providence, Rhode Island.
- CHEN, D., EULENSTEIN, O., FERNÁNDEZ-BACA, D., AND SANDERSON, M. J. 2002a. *Supertrees by Flipping*. Technical Report TR02-01, Department of Computer Science, Iowa State University, 226 Atanasoff Hall, Ames, IA 50011–1040, USA.
- CHEN, D., EULENSTEIN, O., FERNÁNDEZ-BACA, D., AND SANDERSON, M. J. 2002b. Supertrees by flipping. In Ibarra, O. H. and L. Zhang (eds), *Computing and Combinatorics, 8th Annual International Conference, COCOON 2002, Singapore, August 15–17, 2002, Proceedings*, Lecture Notes in Computer Science 2387:391–400. Springer, New York.
- DIAO, L., EULENSTEIN, O., FERNÁNDEZ-BACA, D., AND SANDERSON, M. J. 2003. *Consensus Properties of MRP Supertrees*. Technical Report, Department of Computer Science, Iowa State University, 226 Atanasoff Hall, Ames, IA 50011–1040, USA.
- DOWNEY, R. G. AND FELLOWS, M. R. 1997. *Parameterized Complexity*. Springer, New York.
- DOYLE, J. J. 1992. Gene trees and species trees: molecular systematics as one-character taxonomy. *Systematic Botany* 17:144–163.
- ESTABROOK, G. F., JOHNSON, C., AND McMORRIS, F. R. 1975. An idealized concept of the true cladistic character? *Mathematical Biosciences* 23:263–272.
- EULENSTEIN, O., CHEN, D., BURLEIGH, J. G., FERNÁNDEZ-BACA, D., AND SANDERSON, M. J. In press. Performance of flip-supertrees. *Systematic Biology*.
- FARACH, M., PRZYTYCKA, T., AND THORUP, M. 1995. Agreement of many bounded degree evolutionary trees. *Information Processing Letters* 55:279–301.
- GATESY, J., MATTHEE, C., DESALLE, R., AND HAYAHI, C. 2002. Resolution of a supertree / supermatrix paradox. *Systematic Biology* 51:652–664.
- GATESY, J. AND SPRINGER, M. S. 2004. A critique of matrix representation with parsimony supertrees. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 369–388. Kluwer Academic, Dordrecht, the Netherlands.

- GUSFIELD, D. 1997. *Algorithms on Strings, Trees, and Sequences: Computer Sciences and Computational Biology*. Cambridge University Press, New York.
- HENZINGER, M. R., KING, V., AND WARNOW, T. 1999. Constructing a tree from homeomorphic subtrees, with applications to computational evolutionary biology. *Algorithmica* 24:1–13.
- KEARNEY, P., LI, M., TSANG, J., AND JIANG, T. 1999. Recovering branches on the tree of life: an approximation algorithm. In R. E. Tarjan and T. Warnow (eds), *Symposium on Discrete Algorithms. Proceedings of the Tenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 537–546. Society for Industrial and Applied Mathematics, Philadelphia, PA.
- MAGALLÓN, S. AND SANDERSON, M. J. 2001. Absolute diversification rates in angiosperm clades. *Evolution* 55:1762–1780.
- MATHEWS, S. AND DONOGHUE, M. J. 1999. The root of angiosperm phylogeny inferred from duplicate phytochrome genes. *Science* 286:947–950.
- NATANZON, A., SHAMIR, R., AND SHARAN, R. 2001. Complexity classification of some edge modification problems. *Discrete Applied Mathematics* 113:109–128.
- PAGE, R. D. M. 2002. Modified mincut supertrees. In R. Guigó and D. Gusfield (eds), *Proceedings of the Second International Workshop on Algorithms in Bioinformatics WABI 2002*, pp. 537–552, Springer-Verlag, New York.
- PE’ER, I., SHAMIR, R., AND SHARAN, R. 2000. Incomplete directed perfect phylogeny. In D. Sankoff (ed.), *Proceedings of the Eleventh Symposium on Combinatorial Pattern Matching CPM*, Lecture Notes in Computer Science 1848:143–153. Springer, New York.
- PEETERS, R. 2000. The maximum-edge biclique problem is NP-complete. Research Memorandum 789, Faculty of Economics and Business Administration, Tilburg University.
- PURVIS, A. 1995. A modification to Baum and Ragan’s method for combining phylogenetic trees. *Systematic Biology* 44:251–255.
- QIU, Y. L., LEE, J., BERNASCONI-QUADRINI, F., SOLTIS, D. E., SOLTIS, P. S., ZANIS, M., ZIMMER, E. A., CHEN, Z., SAVOLAINEN, V., AND CHASE, M. W. 1999. The earliest angiosperms: evidence from mitochondrial, plastid, and nuclear genomes. *Nature* 402:404–407.
- RAGAN, M. A. 1992. Phylogenetic inference based on matrix representation of trees. *Molecular Phylogenetics and Evolution* 1:53–58.
- ROSS, H. A. AND RODRIGO, A. G. 2004. An assessment of matrix representation with compatibility in supertree construction. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 35–63. Kluwer Academic, Dordrecht, the Netherlands.
- SANDERSON, M. J. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* 19:301–302.
- SEMPLE, C. AND STEEL, M. 2000. A supertree method for rooted trees. *Discrete Applied Mathematics* 105:147–158.
- SNEDECOR, G. W. AND COCHRAN, W. G. 1995. *Statistical Methods*, 8th ed. Iowa State University Press, Ames, IA.
- SOLTIS, P. S., SOLTIS, D. E., CHASE, M. W., MORT, M. E., ALBACH, D. C., ZANIS, M. J., SAVOLAINEN, V., HAHN, W. H., HOOT, S. B., FAY, M. F., AXTELL, D. C., SWENSON, S. M., PRINCE, L. M., KRESS, W. J., NIXON, K. C., AND FARRIS, J. S. 2000. Angiosperm phylogeny inferred from a combined data set of 18S rDNA, *rbcL*, and *atpB* sequences. *Botanical Journal of Linnean Society* 133:381–461.

- SULLIVAN, J. AND SWOFFORD, D. L. 2001. Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? *Systematic Biology* 50:723–729.
- SWOFFORD, D. L. 1991. When are phylogeny estimates from molecular and morphological data incongruent? In M. M. Miyamoto and J. Cracraft (eds), *Phylogenetic Analysis of DNA Sequences*, pp. 295–333. Oxford University Press, Oxford.
- SWOFFORD, D. L. 2002. *PAUP**. *Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Version 4. Sinauer, Sunderland, Massachusetts.
- YANNAKAKIS, M. 1981. Computing the minimum fill-in is NP-complete. *SIAM Journal on Algebraic and Discrete Methods* 2:77–79.

Chapter 4

EVERYTHING YOU ALWAYS WANTED TO KNOW ABOUT THE AVERAGE CONSENSUS, AND MORE

François-Joseph Lapointe and Claudine Levasseur

Abstract: The average consensus procedure is a method that takes as input a profile of weighted trees and returns a solution that best fits the consensus profile. As an optimization-based approach, it represents one of the few methods available to build consensus trees and supertrees while taking branch lengths into account. The average consensus procedure has been used to address a variety of questions in both the consensus and supertree settings. We present a review of those applications as well as extensions of average consensus trees. The results of new simulations designed to assess the accuracy of average supertrees are also presented and discussed. Finally, we provide recommendations about average consensus trees and suggest future questions to address.

Keywords: average consensus trees, average supertrees, missing distances, phylogenetic accuracy, simulation study, total evidence, weighted trees

1. Introduction

The average consensus procedure (Lapointe and Cucumel, 1997) is a method that takes as input a profile of weighted trees and returns a consensus tree that is in some sense “closest” to the entire profile. Contrary to most standard consensus techniques, the average consensus accounts for branch lengths in the source trees (see also Lapointe, 1998b). Thus, it usually provides consensus trees that are more resolved than those produced by purely topological approaches (e.g., the strict or majority-rule consensus). It is an optimization-based method that returns a solution minimizing a least-

squares criterion that measures the fit of the consensus to the profile of input trees. Alternative criteria have been proposed recently to generalize the average consensus procedure (Levasseur and Lapointe, 2002), and other consensus methods could be characterized in a similar fashion. The original method (Cucumel, 1990), which was designed to combine ultrametric trees or dendograms (i.e., rooted weighted trees in which all leaves are equidistant from the root), has also been extended to allow for the combination of all types of weighted trees, ultrametric or not (Lapointe and Cucumel, 1997). This approach has proven useful to combine phylogenies (see Kirsch *et al.*, 1997; Lapointe *et al.*, 1999), or to validate phylogenetic hypotheses using resampling techniques (Lapointe *et al.*, 1994; Bleiweiss *et al.*, 1994). Furthermore, average consensus supertrees have been published (Lapointe and Kirsch, 2001; Barker, 2002), taking advantages of the mathematical properties of path-length distance matrices associated with weighted trees (see also Lapointe and Cucumel, 1991). Recent studies have compared the accuracy of average consensus trees and total evidence trees in simulation experiments (Levasseur and Lapointe, 2003) as well as with real data (Levasseur and Lapointe, 2001). The behavior of average consensus trees in comparison with consensus methods that ignore branch lengths has also been investigated thoroughly in the past (Lapointe, 1998a; Lapointe *et al.*, 1999).

In the present paper, we will introduce the average consensus procedure and provide a complete review of previous studies published on this subject. First, we will discuss the use of the method in the so-called “consensus setting” (i.e., when all trees in the profile possess identical sets of taxa; *sensu* Bininda-Emonds, 2003), and compare this approach with other consensus methods in the light of the debate on consensus versus total evidence (see de Queiroz *et al.*, 1995). A generalization of optimization-based consensus methods will also be presented as possible extensions to average consensus trees. Second, applications of the average procedure to trees bearing overlapping sets of leaves (the “supertree setting”) will be discussed. We will review the simulation studies conducted to assess the effect of missing distances in phylogenetic studies, with special reference to average supertrees. The problems related to the combination of weighted trees with heterogeneous branch lengths will also be discussed, and new simulations will be presented to illustrate this special case. A general approach for topological distances will be introduced to tackle this problem. Finally, caveats and recommendations will be provided to make the best of the average procedure in a consensus or supertree setting.

2. The average procedure in the consensus setting

Let $S = \{1, 2, \dots, n\}$ be a set of taxa and T_1 and T_2 be two weighted trees defined on S . There exists a one-to-one correspondence between these weighted trees and their associated path-length distance matrices (Hartigan, 1967; Buneman, 1971), where the path length $d(a, b)$ between two taxa a and b is defined as the sum of branch lengths along the path connecting a and b . Thus, it is equivalent to deal with the trees T_1 and T_2 or with their corresponding path-length distance matrices D_1 and D_2 . The least-squares difference Δ between T_1 and T_2 is calculated as (Hartigan, 1967):

$$(1) \quad \Delta(T_1, T_2) = \sum_{a=1}^n \sum_{b=1}^n [d_1(a, b) - d_2(a, b)]^2,$$

where $d_1(a, b)$ and $d_2(a, b)$ are the path-length distances between a and b for T_1 and T_2 , respectively.

Given a profile $P = \{T_1, T_2, \dots, T_k\}$ of k weighted trees defined on a common set of taxa S , the average consensus tree T_c is defined as the weighted tree that minimizes the following function Q (Vichi, 1993; Lapointe and Cucumel, 1997):

$$(2) \quad Q = \sum_{i=1}^k \Delta(T_c, T_i).$$

With no loss of generality, one could consider this consensus procedure as one involving a profile of scaled trees represented by their path-length distance matrices taking their values in the interval $[0, 1]$. For simplicity, the average consensus tree T_c is computed usually in a stepwise fashion. First, a distance matrix \bar{D} is obtained by computing the average of the k path-length distance matrices associated with the trees of P , such that:

$$(3) \quad \bar{d}(a, b) = \frac{1}{k} \sum_{i=1}^k d_i(a, b),$$

where $\bar{d}(a, b)$ are the average distances and $d_i(a, b)$ are the path-length distances associated with the tree T_i . In those cases where all trees of P are identical topologically, \bar{D} represents a path-length distance matrix corresponding to a unique consensus tree T_c with average branch lengths. In the majority of other cases, however, the average consensus T_c is obtained as the solution that minimizes the function Q' :

$$(4) \quad Q' = \sum_{a=1}^n \sum_{b=1}^n [d_c(a, b) - \bar{d}(a, b)]^2,$$

where $d_c(a, b)$ are the fitted path-length distances of T_c . Minimizing Q' (Eq. 4) is equivalent to minimizing Q (Eq. 2) (for a proof, see Lapointe and Cucumel, 1997). Thus, every least-squares algorithm (Cavalli-Sforza and Edwards, 1967; Fitch and Margoliash, 1967; De Soete, 1983; Makarenkov and Leclerc, 1999) can be used to compute an approximation of the average consensus tree T_c from the average distance matrix \bar{D} .

2.1 A family of optimization-based consensus methods

The average procedure is not the only optimization-based consensus method. As a matter of fact, the least-squares criterion (Eq. 4) is only one of numerous optimality criteria that can be used to define consensus functions with distinct properties. A family of average consensus procedures has been introduced by Levasseur and Lapointe (2002), including the mean consensus (here referred to as the average consensus) as a special case of the more general approach. The methods of this family can differ with respect to 1) the types of trees represented by distance matrices, 2) the way these matrices are combined to yield a single matrix, and 3) the optimization criterion selected to compute the consensus tree. For instance, let Δ define the absolute difference between two weighted trees:

$$(5) \quad \Delta|T_1, T_2| = \sum_{a=1}^n \sum_{b=1}^n [d_1(a, b) - d_2(a, b)]^2,$$

where $d_1(a, b)$ and $d_2(a, b)$ are the path-length distances for T_1 and T_2 , respectively. Given a profile $P = \{T_1, T_2, \dots, T_k\}$ of k weighted trees defined on a common set of taxa S , the median consensus weighted tree T_c (not to be confounded with the median consensus for n -trees; Barthélemy and McMorris, 1986) is defined as the solution that minimizes the following function L :

$$(6) \quad L = \sum_{i=1}^n \Delta|T_c, T_i|.$$

Interestingly, the median consensus tree can be obtained in a stepwise fashion, just like any average consensus tree. The first step involves the

computation of a median distance matrix D_{med} from the path-length distance matrices representing the trees of P . When the number of trees k is odd, the median distances are given by the $(k + 1) / 2$ -th ordered $d(a, b)$ distances in the path-length distance matrices. When k is even, the median distances are computed as the mean of the $k / 2$ -th and $(k / 2 + 1)$ -th ordered $d(a, b)$ values in the path-length distance matrices. A sum-of-absolute differences algorithm (Smith, 2001) is applied to the median matrix D_{med} to obtain the median consensus tree T_c that minimizes the following function L' :

$$(7) \quad L' = \sum_{a=1}^n \sum_{b=1}^n |d_c(a, b) - d_{\text{med}}(a, b)|,$$

where $d_c(a, b)$ are the fitted path-length distances of T_c and $d_{\text{med}}(a, b)$ are the distances in the median matrix D_{med} .

Other consensus functions can be defined using different optimality measures, such as the mixed mean-median consensus that combines two minimization criteria (Eqs 2 and 6) as follows (Mirkin and Roberts, 1993):

$$(8) \quad R = \alpha Q(1 - \alpha) L,$$

where α can take all values on the interval $[0, 1]$. Using this model, the median consensus is obtained when $\alpha = 0$, and the mean (average) consensus is obtained when $\alpha = 1$.

Depending on the type of trees of P (e.g., weighted or unweighted, rooted or unrooted), different consensus solutions could also be derived. For example, an ultrametric constraint could be imposed to compute the consensus of a profile of dendograms (ultrametric weighted trees), as proposed originally by Cucumel (1990), and consensus networks could also be obtained by applying a minimization criterion to find the split-decomposition graph (Bandelt and Dress, 1992; Dress *et al.*, 1996; Huson, 1998) that best fits a profile of weighted trees. All these procedures take advantage of the one-to-one correspondence between weighted trees and their associated path-length distance matrices (Hartigan, 1967). Different optimality criteria are available for other representations of trees, however. The spectral consensus (Lapointe, 1998b) is an optimization-based method based on tree spectra (Hendy and Penny, 1993) that uses the average spectrum of weighted trees to derive a consensus solution closest to the input profile P . Similarly, the well-known matrix representation with parsimony supertree method (MRP; Baum, 1992; Ragan, 1992) can also be defined in the consensus setting as an optimization-based approach that finds the tree

that best fits the profile of input trees (with or without branch lengths) using a parsimony algorithm (Baum and Ragan, 1993; Bryant, 2003).

2.2 The total evidence – consensus debate

Several papers have been published on the use of consensus methods to combine the results of separate phylogenetic analyses (for reviews, see de Queiroz *et al.*, 1995; Huelsenbeck *et al.*, 1996). In general, it appears that standard consensus techniques that ignore branch lengths are less accurate than a total-evidence approach that combines all data before phylogenetic analysis (Kluge 1989; Barrett *et al.*, 1991). By contrast, Lapointe (1998a) illustrated that average consensus trees could be as resolved as, or identical to, those obtained by total evidence. Lapointe *et al.* (1999) showed further that such congruent results are to be expected when data and trees are treated coherently as distance matrices (for a generalization of these results in the supertree setting, see also Kirsch *et al.*, 1997; Lapointe and Kirsch, 2001).

To assess the relative performance of average consensus trees in comparison with total-evidence trees, Levasseur and Lapointe (2001) reanalyzed 15 published data sets using a distance-based approach. Their results showed clearly that average consensus trees are equivalent to total-evidence trees, and that both methods outperform standard consensus techniques such as strict (Sokal and Rohlf, 1981) and majority-rule consensus (Margush and McMorris, 1981). They also showed that all discrepancies were caused by poorly supported nodes (i.e., those with low bootstrap values). To corroborate these findings, a simulation study was set up by Levasseur and Lapointe (2003) to measure phylogenetic accuracy and assess the congruence among the competing approaches when applied jointly or separately. These simulations provided similar results to those obtained with real data (Levasseur and Lapointe, 2001): average consensus trees can be as good as total evidence trees, or, in other words, combining trees or data often produce identical results when a consensus method that accounts for branch lengths is selected.

In the light of the previous studies, it clearly appears that the average consensus represents a powerful tool to combine the results of independent phylogenetic analyses. Indeed, our results showed that average consensus trees are more accurate than total evidence (Levasseur and Lapointe, 2003), especially when the data partitions are not homogeneous on the basis of statistical heterogeneity tests (Farris *et al.*, 1995). We believe firmly that the debate over the use of consensus in phylogenetics is biased with respect to the consensus method selected. We propose to rely not on a single method, but to use separate and combined analyses together in a so-called global congruence approach (*sensu* Lapointe *et al.*, 1999). Indeed, the simulation

results revealed that phylogenetic accuracy was increased when average consensus trees were congruent (i.e., identical topologically) with the corresponding total evidence trees (Levasseur and Lapointe, 2003).

3. The average procedure in the supertree setting

Let $P = \{T_1, T_2, \dots, T_k\}$ be a profile of k weighted trees T_i defined on different sets of taxa S_i , such that $S = S_1 \cup S_2 \dots \cup S_k$. The average consensus supertree T_c is a weighted tree, defined on S , that is closest to P , using a least-squares criterion (Lapointe and Cucumel, 1997). The average procedure in the consensus setting (Eq. 2) is a special case of the supertree setting, when all trees of P are defined on the same set of taxa. Computation of the average supertree is therefore very similar to the computation of the average consensus tree. The major difference lies in the calculation of a weighted average distance matrix \bar{D}^* that accounts for the fact that some taxa might not be represented in all trees of P . The matrix \bar{D}^* is computed as:

$$(9) \quad \bar{d}^*(a, b) = \frac{1}{w(a, b)} \sum_{i=1}^k d_i(a, b),$$

where $\bar{d}^*(a, b)$ are the average distances, $d_i(a, b)$ are the path-length distances associated with the tree T_i (taken to be 0 if T_i does not contain both a and b), and the $w(a, b)$ equals the number of trees in P that contain both the taxa a and b .

Two distinct situations need to be considered here, depending on whether \bar{D}^* is complete or incomplete. The average matrix will be complete if all pairwise distances of \bar{D}^* are defined, implying that every pair of taxa a and b is contained in at least one tree of P . (In the consensus setting, all pairs of taxa a and b are found in all k trees of P .) In such cases, the minimization step of the algorithm takes as input the average distance matrix \bar{D}^* and returns the weighted supertree T_c that is closest to P in a least-squares sense. When \bar{D}^* is incomplete, some of the pairwise distances are undefined, and a different approach is required.

3.1 Incomplete distance matrices and supertrees

To solve the problems encountered in the combination of weighted trees bearing overlapping sets of taxa, several methods have been proposed to estimate missing distances before phylogenetic analysis (for a review, see

Landry and Lapointe, 1997). De Soete (1984a, b) and Lapointe and Kirsch (1995) have relied on the mathematical properties of ultrametric distances (Hartigan, 1967) to deal with incomplete matrices. Ultrametric estimates are obtained by applying the following equation recursively to every missing distance in the matrix:

$$(10) \quad d(a, b) = \max[d(a, c); d(b, c)],$$

where $d(a, b)$ is missing and the other two distances are known; the minimum distance estimated for all triplets $\{a, b, c\}$ is returned as the final estimate of $d(a, b)$.

Whereas the ultrametric approach has proven to be very accurate for estimating missing distances in ultrametric matrices representing dendograms (De Soete, 1984a), this algorithm is far from perfect for estimating missing values in additive path-length distance matrices associated with non-ultrametric weighted trees (De Soete, 1984b). In such cases, the four-point condition (Buneman, 1974) can be used to obtain better estimates by applying the following equation recursively to every missing distance in the matrix (Landry *et al.*, 1996):

$$(11) \quad d(a, b) = \max[d(a, d) + d(b, c); d(a, c) + d(b, d)] - d(c, d),$$

where $d(a, b)$ is missing and the other five distances are known; the minimum distance estimated for all quartets $\{a, b, c, d\}$ is returned as the final estimate of $d(a, b)$.

Many simulation studies have addressed the performance of additive and ultrametric estimation procedures (Landry *et al.*, 1996; Landry and Lapointe, 1997; Levasseur *et al.*, 2000), and these methods have been used successfully in the past to build phylogenies from incomplete distance matrices (e.g., Kirsch *et al.*, 1997). More recently, Lapointe and Landry (2001) proposed an estimation method that relies on the decomposition of additive distance matrices into an ultrametric and a star component (see Lapointe and Legendre, 1992). Instead of applying the four-point condition (Eq. 11), the ultrametric part of the matrix is estimated using the ultrametric property (Eq. 10), after which the star component is added to it to obtain a complete additive distance matrix. The approach based on triplets is faster than using quartets, while satisfying the properties of additive distance matrices.

The properties of path-length distance matrices guarantee that missing cells can be estimated accurately in matrices that satisfy the four-point condition (Eq. 11), at least in theory. In practice, however, many factors can affect such estimations (see Landry and Lapointe, 1997). For one, the

number of missing cells in a distance matrix has a direct impact on accuracy because incorrect estimates can be used to estimate other missing cells in turn. This type of chain reaction is particularly important in trees with short internodes, where a small error in the estimation of a missing distance can have a significant effect in terms of topology. A similar problem is related to the filling order of missing cells. Indeed, simulations have shown that different permutations of a distance matrix can lead to different estimates because the order in which the missing cells are filled is not the same (Landry and Lapointe, 1997). Finally, the most important problem with incomplete path-length distance matrices is the estimation of a missing distance between sister taxa (Lapointe and Kirsch, 1995; Landry and Lapointe, 1997). In such cases, the actual distance will never be estimated correctly by either the ultrametric (Eq. 10) or additive (Eq. 11) methods, or any other estimation method for that matter. At best, the sister pair will collapse to a trichotomy with the next most closely related taxa (Lapointe and Kirsch, 1995).

In spite of these problems, the estimation of missing distances represents an interesting solution that allows the computation of average supertrees from an incomplete matrix of average distances \bar{D}^* . This solution is not the only option available, however. Simple methods have been proposed to reconstruct weighted trees directly from incomplete sets of distances (Hein, 1989; Guénoche and Grandcolas, 1999), whereas Makarenkov and Leclerc (1999) have published a weighted-least-squares algorithm recently that also applies to incomplete distance matrices. This approach can be used to compute the average supertree T_c by minimizing the following function Q^* :

$$(12) \quad Q^* = \sum_{a=1}^n \sum_{b=1}^n \frac{1}{w(a,b)} [d_c(a,b) - \bar{d}^*(a,b)]^2,$$

where $\bar{d}^*(a,b)$ are the average distances, $d_c(a,b)$ are the fitted path-length distances of T_c , and $w(a,b)$ are a set of binary weights that are set to zero for missing cells and to one for all other distances of \bar{D}^* . Notice that it becomes equivalent to minimize Q' (Eq. 4) or Q^* (Eq. 12) when the average distance matrix \bar{D}^* is complete.

It is not clear whether the average supertrees derived from incomplete average matrices satisfy the same properties as the trees obtained in the consensus setting. Simulations have also shown that the indirect estimation methods (ultrametric or additive) and the direct algorithmic approaches can produce different but somewhat equivocal results. Additive estimates are more accurate in cases where few distances are missing, whereas ultrametric estimates do better as the number of missing distances increases (Landry *et*

al., 1996; Landry and Lapointe, 1997). Contrary to the findings of Makarenkov (2002), the weighted least-squares approach (Eq. 12) is not as good as either indirect method in simulations involving randomly generated matrices with missing distances (Levasseur *et al.*, 2003). Further simulations are needed badly to examine the behavior of the average procedure in the supertree setting, however.

3.2 Applications of average supertrees

Besides the classical use of supertree methods to synthesize the results of independent analyses, average supertrees can also be useful to assess the stability of weighted trees. It is well known that taxonomic sampling can have a major impact on phylogenetic trees (Lecointre *et al.*, 1993). Lapointe *et al.* (1994) have taken advantage of the average consensus to assess this effect using a taxonomic jackknife procedure (Lanyon, 1985). To do so, replicate matrices are created by deleting a certain number of taxa at random from a complete distance matrix. Phylogenetic trees are then derived from each replicate matrix, and the average procedure is used to combine the trees. The result is a supertree defined on the complete set of taxa. If the data are consistent internally, all trees will be compatible topologically, and the average supertree will be congruent with the phylogeny derived from the complete set of objects. Any discrepancies indicate the presence of unstable taxa. This method has been used mostly to validate phylogenies constructed from distance data (e.g., Bleiweiss *et al.*, 1994; Kirsch *et al.*, 1995; Campeau-Péloquin *et al.*, 2001), but the same approach can be used with other methods of phylogenetic analysis (e.g., maximum likelihood) so long as weighted trees are combined. In this specific context, the average consensus could be considered as an alternative to quartet puzzling (Strimmer and von Haeseler, 1996) to assemble trees while taking their branch lengths into account.

Supertrees are often used in comparative analyses to study the evolution of specific life-history traits in a phylogenetic framework (see Bininda-Emonds *et al.*, 2002; Gittleman *et al.*, 2004). The comparative method requires complete phylogenies (but see Losos, 1994), and branch lengths are often used for the computation of corrected data (Felsenstein, 1985). The average procedure was used by Kirsch *et al.* (1997) to assemble what is currently the largest published phylogeny of marsupials (see also Lapointe and Kirsch, 2001). That supertree was then used by Johnson (1998) to test hypotheses of species extinction using a comparative approach. Similarly, Barker (2002) constructed supertrees using MRP and average consensus methods to assess the utility of phylogenetic diversity measures (Faith, 1992).

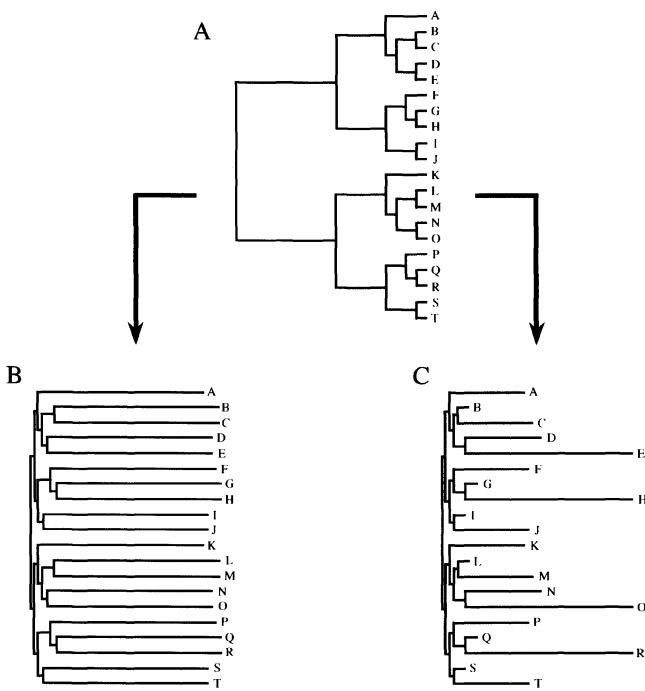


Figure 1. Model topology (A) with different branch lengths and with (B) slow (0.25) and (C) rapid (1.75) evolutionary rates of change along the branches.

3.3 Assessing the accuracy of average supertrees

A simulation study was undertaken to assess the accuracy of the average consensus in the supertree setting using a model tree with 20 taxa (Figure 1A) inspired by Kumar (1996). To make things simple, we restricted ourselves to cases involving the combination of only two weighted trees, representing subtrees of the model supertree. Different parameters were investigated in the simulations: the relative sizes of the subtrees (identical or different), the size of the overlap between subtrees (small or large), and the degree of heterogeneity of the data evolved on the model tree (homogeneous or heterogeneous). To simulate heterogeneous data sets, DNA sequences of 2500 bp were evolved independently on two model trees with the same topology, but with different branch lengths and using different rates of evolution (Figures 1B, C). Homogeneous data sets were simulated by evolving DNA sequences of 2500 bp on the same tree (Figure 1C). The heterogeneity of the data sets was assessed with the incongruence length

difference test (ILD; Farris *et al.*, 1995). Four situations were simulated for heterogeneous and homogeneous data sets: 1) subtrees of the same size (13 taxa) with a small overlap (six taxa), 2) subtrees of the same size (15 taxa) with a large overlap (10 taxa), 3) subtrees of different sizes (10 and 16 taxa) with a small overlap (six taxa), and 4) subtrees of different sizes (13 and 17 taxa) with a large overlap (10 taxa). For each case, 1000 replicates were simulated. All sequences were generated with the Seq-Gen program (Rambaut and Grassly, 1997) using a Jukes-Cantor model of evolution (Jukes and Cantor, 1969).

Distance matrices were computed from the DNA sequences using a Jukes-Cantor correction, and trees were estimated from these distances using an unweighted least-squares method (Cavalli-Sforza and Edwards, 1967) in PAUP* (Swofford, 2002). Three different standardization techniques were employed to correct for differences in branch lengths caused by the relative sizes of the subtrees and the heterogeneous rates of evolution. Namely, the distance values in each matrix were scaled either by 1) dividing all distances by the maximum distance in the entire matrix, 2) dividing all distances by the maximum distance in the common part of the matrix representing the overlap of the two subtrees, or 3) by multiplying the distances in the first matrix, such as to maximize the fit to the second matrix. These corrected path-length distance matrices associated with the corresponding subtrees were combined to compute an average matrix defined on the whole set of taxa. The missing distances in the non-overlapping part of the matrix were estimated with the additive procedure (Eq. 11). The average supertree was then obtained by applying a least-squares algorithm to the matrix.

Accuracy was measured by comparing each average supertree to the model tree (Figure 1A). To do so, the strict consensus of these two trees was computed and the consensus fork index (*CFI*; Colless, 1980) was used to quantify topological agreement. The *CFI* measures the relative number of resolved clades in the strict consensus tree. Its maximum value of one indicates total congruence (i.e., the average supertree and the model tree are identical topologically and their strict consensus is fully resolved), whereas a value of zero is indicative of total incongruence (i.e., the average supertree and the model tree are incompatible topologically and their strict consensus is a bush). The mean *CFI* values of the 1000 replicates and their standard deviations are reported in Table 1 for the four situations considered in the simulations. Because the results obtained with the three different standardization techniques were very similar, only those corresponding to the third method are presented. In the case of homogeneous data sets, the mean *CFI* values range from 0.720 to 0.879, and indicate that accuracy is improved when the overlap is large and the number of missing cells is small. By contrast, the values obtained for heterogeneous data sets are much worse

Table 1. Mean consensus fork index values obtained in the four situations considered in the simulations for homogeneous and heterogeneous data sets. Actual branch lengths were used to compute average supertrees. The standard deviations are given in parentheses. All simulations were based on 1000 replicates.

	Size: subtree 1 (overlap) subtree 2			
	13(6)13	10(6)16	15(10)15	13(10)17
Number of missing distances	49	40	25	21
Homogeneous data sets	0.720 (0.116)	0.771 (0.109)	0.868 (0.089)	0.879 (0.087)
Heterogeneous data sets	0.147 (0.096)	0.200 (0.099)	0.081 (0.073)	0.071 (0.064)

(ranging from 0.071 to 0.200), and do not follow the same trend as homogeneous data sets. The differences in branch lengths and rates of evolution clearly affect the accuracy of average supertrees in such situations, regardless of the scaling method selected. To avoid this problem, the same simulations were repeated by setting all branch lengths to one before combining the subtrees. The corresponding branch-distance matrices were thus used to compute an average supertree defined on topological relationships alone, almost like MRP, but using a different matrix representation. The results of these simulations are presented in Table 2. Following this modification, the mean *CFI* values for homogeneous data sets decreased slightly, but the corresponding values for heterogeneous data sets increased dramatically compared with the results of the first series of simulations (see Table 1).

The combination of branch-distance matrices represents a promising extension of the average procedure that deserves further exploration. Given our findings, it would seem sensible to ignore branch length information in building the average supertree because the penalty in doing so when the data sets are homogeneous is slight, but the benefit in doing so when they are heterogeneous is great. In other words, the major strength of the average procedure can become a weakness when source trees with heterogeneous branch lengths are combined. Similarly, combining trees with branch lengths with others that do not have branch lengths will clearly affect the resulting supertree. In such cases, it is always preferable to ignore branch lengths altogether when building average supertrees. For the same reason, we do not recommend combining branch-distance matrices with path-length distance matrices. When branch lengths are available, however, standardization methods must be used to scale path-length distances in such a way that they become comparable. We have addressed this problem in our simulations, but further studies would be required to examine the accuracy of average supertrees fully, with or without branch lengths (see also Bininda-Emonds

Table 2. Mean consensus fork index values obtained in the four situations considered in the simulations for homogeneous and heterogeneous data sets. All branch lengths were set to one to compute average supertrees. The standard deviations are given in parentheses. All simulations were based on 1000 replicates.

	Size: subtree 1 (overlap) subtree 2			
	13(6)13	10(6)16	15(10)15	13(10)17
Number of missing distances	49	40	25	21
Homogeneous data sets	0.703 (0.108)	0.752 (0.099)	0.863 (0.084)	0.878 (0.082)
Heterogeneous data sets	0.615 (0.126)	0.691 (0.117)	0.777 (0.109)	0.763 (0.111)

and Sanderson, 2001). Other ways of scaling path-length distance matrices also need to be investigated when combining more than two trees of varying sizes. Finally, the relative performance of average supertrees with respect to other supertree methods (Gordon, 1986; Baum, 1992; Lanyon 1993; Semple and Steel, 2000; Goloboff and Pol, 2002) must be addressed, as well as the relationship between average supertrees and total evidence trees derived from incomplete data sets.

4. A final word on average consensus trees and supertrees

The average consensus procedure is one of the few methods available to derive consensus trees and supertrees with branch lengths (see also Stinebrickner 1984, Lefkovich, 1985; Brossier 1990, Lapointe, 1998b; Bryant, 2003). As an optimization-based approach, it differs from other methods that can be defined by simple constructive algorithms (Barthélemy *et al.*, 1986). The computation of an average consensus tree is an NP-hard problem (Barthélemy and Brucker, 2000) and heuristics must be used to minimize the least-squares criterion. Thus, optimal solutions are not always guaranteed. Furthermore, the average consensus tree might not always be unique (see Wilkinson *et al.*, 2003). In such cases, a topological consensus of the multiple solutions can be used at the cost of losing information about branch lengths. The average consensus procedure also differs from consensus and supertree methods that are characterized using an axiomatic approach (McMorris, 1985). For example, the co-Pareto axiom can be violated by average consensus trees, implying that clades appearing in the consensus solution might not be found in the input trees (see Levasseur and Lapointe, 2001). It is not clear whether average consensus trees satisfy any other desirable properties of standard consensus methods based on topology.

In his classification of consensus techniques, Bryant (2003) defined the average consensus as an outlier without any relationship to other consensus tree methods. As an indirect consensus method (*sensu* Wilkinson *et al.*, 2001), it is related to MRP in the consensus setting, however (Bryant, 2003). When branch distances are used instead of path-length distances, average consensus and MRP are equivalent methods, only using distinct optimality criteria (Lapointe *et al.*, 2003). Analytical and empirical studies are badly needed to investigate this relationship in the supertree setting.

Acknowledgements

We would like to thank Mike Steel for his constructive comments on this paper, and Olaf Bininda-Emonds for his editorial work. This research was funded by NSERC grant OGP0155251 to François-Joseph Lapointe and by NSERC scholarship to Claudine Levasseur.

References

- BANDELT, H. J. AND DRESS, A. W. 1992. Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Molecular Phylogenetics and Evolution* 1:242–252.
- BARKER, G. M. 2002. Phylogenetic diversity: a quantitative framework for measurement of priority and achievement in biodiversity conservation. *Biological Journal of the Linnean Society* 76:165–194.
- BARRETT, M., DONOGHUE, M. J., AND SOBER, E. 1991. Against consensus. *Systematic Zoology* 40:486–493.
- BARTHÉLEMY, J. P. AND BRUCKER, F. 2000. Average consensus in numerical taxonomy and some generalizations. In W. Gaul, O. Opitz, and M. Schader (eds), *Data Analysis: Scientific Modeling and Practical Application*, pp. 95–104. Springer-Verlag, Berlin.
- BARTHÉLEMY, J. P., LECLERC, B., AND MONJARDET, B. 1986. On the use of ordered sets in problems of comparison and consensus classifications. *Journal of Classification* 3:187–224.
- BARTHÉLEMY, J.-P. AND McMORRIS, F. R. 1986. The median procedure for n -trees. *Journal of Classification* 3:329–334.
- BAUM, B. R. 1992. Combining trees as a way of combining data for phylogenetic inference, and the desirability of combining gene trees. *Taxon* 41:3–10.
- BAUM, B. R. AND RAGAN, M. A. 1993. Reply to A. G. Rodrigo's “A comment on Baum's method for combining phylogenetic trees”. *Taxon* 42:637–640.
- BININDA-EMONDS, O. R. P. 2003. MRP supertree construction in the consensus setting. In M. Janowitz, F.-J. Lapointe, F. R. McMorris, B. Mirkin, and F. S. Roberts (eds), *Bioconsensus*, pp. 231–242. American Mathematical Society, Providence, Rhode Island.
- BININDA-EMONDS, O. R. P., GITTLEMAN, J. L., AND STEEL, M. A. 2002. The (super)tree of life: procedures, problems and prospects. *Annual Review of Ecology and Systematics* 33:265–289.

- BININDA-EMONDS, O. R. P. AND SANDERSON, M. J. 2001. Assessment of the accuracy of matrix representation with parsimony analysis supertree construction. *Systematic Biology* 50:565–579.
- BLEIWEISS, R., KIRSCH, J. A. W., AND LAPOINTE, F.-J. 1994. DNA-DNA hybridization-based phylogeny of higher nonpasserines: reevaluating a key portion of the avian family tree. *Molecular Phylogenetics and Evolution* 3:248–255.
- BROSSIER, G. 1990. Piecewise hierarchical clustering. *Journal of Classification* 7:197–216.
- BRYANT, D. 2003. A classification of consensus methods for phylogenetics. In M. Janowitz, F.-J. Lapointe, F. R. McMorris, B. Mirkin, and F. S. Roberts (eds), *Bioconsensus*, pp. 163–184. American Mathematical Society, Providence, Rhode Island.
- BUNEMAN, P. 1971. The recovery of trees from measures of dissimilarity. In F. R. Hodson, Kendall, D. G., and Tautu, P. (eds), *Mathematics in Archeological and Historical Sciences*, pp. 387–395. Edinburgh University Press, Edinburgh.
- BUNEMAN, P. 1974. A note on the metric properties of trees. *Journal of Combinatorial Theory (B)* 17:48–50.
- CAMPEAU-PÉLOQUIN, A., KIRSCH, J. A. W., ELDRIDGE, M. D. B., AND LAPOINTE, F.-J. 2001. Phylogeny of the rock-wallabies, *Petrogale* (Marsupialia: Macropodidae) based on DNA/DNA hybridisation. *Australian Journal of Zoology* 49:463–486.
- CAVALLI-SFORZA, L. L. AND EDWARDS, A. W. F. 1967. Phylogenetic analysis: models and estimation procedures. *American Journal of Human Genetics* 19:233–257.
- COLLESS, D. H. 1980. Congruence between morphometric and allozyme data for *Menidia* species: a reappraisal. *Systematic Zoology* 29:288–299.
- CUCUMEL, G. 1990. Construction d'une hiérarchie consensus à l'aide d'une ultramétrique centrale. In *Recueil des Textes des Présentations du Colloque sur les Méthodes et Domaines d'Application de la Statistique 1990*, pp. 235–243. Bureau de la Statistique du Québec, Québec.
- DE QUEIROZ, A., DONOGHUE, M. J., AND KIM, J. 1995. Separate versus combined analysis of phylogenetic evidence. *Annual Review of Ecology and Systematics* 26:657–681.
- DE SOETE, G. 1983. A least squares algorithm for fitting additive trees to proximity data. *Psychometrika* 48:621–626.
- DE SOETE, G. 1984a. Ultrametric tree representations of incomplete dissimilarity data. *Journal of Classification* 1:235–242.
- DE SOETE, G. 1984b. Additive-tree representations of incomplete dissimilarity data. *Quality and Quantity* 18:387–393.
- DRESS A., HUSON, D., AND MOULTON V. 1996. Analyzing and visualizing sequence and distance data using SplitsTree. *Discrete Applied Mathematics* 71:95–109.
- FAITH, D. P. 1992. Conservation evaluation and phylogenetic diversity. *Biological Conservation* 61:1–10.
- FARRIS, J. S., KÄLLERSJÖ, A. G., KLUGE, A. G., AND BULT, C. 1995. Testing significance of incongruence. *Cladistics* 10:315–319.
- FITCH, W. M. AND MARGOLIASH, E. 1967. Construction of phylogenetic trees. *Science* 155:279–284.
- FELSENSTEIN, J. 1985. Phylogenies and the comparative method. *The American Naturalist* 125:1–15.
- GITTLEMAN, J. L., JONES, K. E., AND PRICE, S. A. 2004. Supertrees: using complete phylogenies in comparative biology. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 439–460. Kluwer Academic, Dordrecht, the Netherlands.
- GOLOBOFF, P. A. AND POL, D. 2002. Semi-strict supertrees. *Cladistics* 18:514–525.

- GORDON, A. D. 1986. Consensus supertrees: the synthesis of rooted trees containing overlapping sets of labeled leaves. *Journal of Classification* 3:335–348.
- GUÉNOCHE, A. AND GRANDCOLAS, S. 1999. Approximations par arbre d'une distance partielle. *Mathématique, Informatique et Sciences Humaines* 146:51–64.
- HARTIGAN, J. A. 1967. Representation of similarity matrices by trees. *Journal of the American Statistical Association* 62:1140–1158.
- HEIN, J. 1989. A tree reconstruction method that is economical in the number of pairwise comparison used. *Molecular Biology and Evolution* 6:669–684.
- HENDY, M. D. AND PENNY, D. 1993. Spectral analysis of phylogenetic data. *Journal of Classification* 10:5–24.
- HUELSENBECK, J. P., BULL, J. J., AND CUNNINGHAM, C. W. 1996. Combining data in phylogenetic analysis. *Trends in Ecology and Evolution* 11:152–158.
- HUSON, D. H. 1998. SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* 14:68–73.
- JOHNSON, C. N. 1998. Species extinction and the relationship between distribution and abundance. *Nature* 394:272–274.
- JUKES, T. H. AND CANTOR, C. R. 1969. Evolution of protein molecules. In H. N Munro (ed.), *Mammalian Protein Metabolism*, pp. 21–132. Academic Press, New York.
- KIRSCH, J. A. W., LAPOINTE, F.-J., AND FOESTE, A. 1995. Resolution of portions of the kangaroo phylogeny (Marsupialia: Macropodidae) using DNA hybridization. *Biological Journal of the Linnean Society* 55:309–328.
- KIRSCH, J. A. W., LAPOINTE, F.-J., AND SPRINGER, M. S. 1997. DNA-hybridisation studies of marsupials and their implications for metatherian classification. *Australian Journal of Zoology* 45:211–280.
- KLUGE, A. G. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). *Systematic Biology* 38:7–25.
- KUMAR, S. 1996. A stepwise algorithm for finding minimum evolution trees. *Molecular Biology and Evolution* 13:584–593.
- LANDRY, P.-A. AND LAPOINTE, F.-J. 1997. Estimation of missing distances in path-length matrices: problems and solutions. In B. Mirkin, F. R. McMorris, F. S. Roberts, and A. Rzhetsky (eds), *Mathematical Hierarchies and Biology*, pp. 209–218. American Mathematical Society, Providence, Rhode Island.
- LANDRY, P.-A., LAPOINTE F.-J., AND KIRSCH, J. A. W. 1996. Estimating phylogenies from lacunose distance matrices: additive is superior to ultrametric estimation. *Molecular Biology and Evolution* 13:818–823.
- LANYON, S. 1985. Detecting internal inconsistencies in distance data. *Systematic Zoology* 34:397–403.
- LANYON, S. 1993. Phylogenetic frameworks: towards a firmer foundation for the comparative approach. *Biological Journal of the Linnean Society* 49:45–61.
- LAPOINTE, F.-J. 1998a. How to validate phylogenetic trees? A stepwise procedure. In C. Hayashi, H. H. Bock, K. Yajima, Y. Tanaka, N. Oshumi, and Y. Baba (eds), *Data Science, Classification, and Related Methods: Studies in Classification, Data Analysis, and Knowledge Optimization*, pp. 71–88. Springer-Verlag, Tokyo.
- LAPOINTE, F.-J. 1998b. For consensus (with branch lengths). In A. Rizzi, M. Vichi, and H.-H. Bock (eds), *Advances in Data Science and Classification*, pp. 73–80. Springer-Verlag, Heidelberg.
- LAPOINTE, F.-J. AND CUCUMEL, G. 1991. Le super-dendrogramme ou la combinaison de matrices ultramétriques partiellement disjointes. In *Recueil des Textes des Présentations*

- du Colloque sur les Méthodes et Domaines d'Application de la Statistique 1991*, pp. 145–151. Bureau de la Statistique du Québec, Québec.
- LAPOINTE, F.-J. AND CUCUMEL, G. 1997. The average consensus procedure: combination of weighted trees containing identical or overlapping sets of objects. *Systematic Biology* 46:306–312.
- LAPOINTE, F.-J. AND KIRSCH, J. A. W. 1995. Estimating phylogenies from lacunose distance matrices, with special reference to DNA hybridization data. *Molecular Biology and Evolution* 12:266–284.
- LAPOINTE, F.-J. AND KIRSCH, J. A. W. 2001. Construction and verification of a large phylogeny of marsupials. *Australian Mammalogy* 23:9–22.
- LAPOINTE, F.-J., KIRSCH, J. A. W., AND BLEIWEISS, R. 1994. Jackknifing of weighted trees: validation of phylogenies reconstructed from distances matrices. *Molecular Phylogenetics and Evolution* 3:256–267.
- LAPOINTE, F.-J., KIRSCH, J. A. W., AND HUTCHEON, J. M. 1999. Total evidence, consensus, and bat phylogeny: a distance based approach. *Molecular Phylogenetics and Evolution* 11:55–66.
- LAPOINTE, F.-J. AND LANDRY, P.-A. 2001. A fast procedure for estimating missing distances in incomplete matrices prior to phylogenetic analysis. In N. El-Mabrouk, T. Lengauer, and D. Sankoff (eds), *Currents in Computational Molecular Biology*, pp. 189–190. Publications CRM, Montréal.
- LAPOINTE, F.-J. AND LEGENDRE, P. 1992. A statistical framework to test the consensus among additive trees (cladograms). *Systematic Biology* 41:158–171.
- LAPOINTE, F.-J., WILKINSON, M., AND BRYANT, D. 2003. Matrix representations with parsimony or with distances: two sides of the same coin? *Systematic Biology* 52:865–868.
- LECOINTRE, G. H., PHILIPPE, H., VÂN LÊ, H. L., AND LE GUYADER, H. 1993. Species sampling has a major impact on phylogenetic inference. *Molecular Phylogenetics and Evolution* 2:205–224.
- LEFKOVITCH, L. P. 1985. Euclidean consensus dendograms and other classification structures. *Mathematical Biosciences* 74:1–15.
- LEVASSEUR, C., LANDRY, P.-A., AND LAPOINTE, F.-J. 2000. Estimating trees from incomplete distance matrices: a comparison of two methods. In H. A. L. Kiers, J.-P. Rasson, P. J. F. Groenen, and M. Schader (eds), *Data Analysis, Classification, and Related Methods*, pp. 149–154. Springer-Verlag, Berlin.
- LEVASSEUR, C., LANDRY, P.-A., MAKARENKO, V., KIRSCH, J. A. W., AND LAPOINTE, F.-J. 2003. Incomplete distance matrices, supertrees and bat phylogeny. *Molecular Phylogenetics and Evolution* 27:239–246.
- LEVASSEUR, C. AND LAPOINTE, F.-J. 2001. War and peace in phylogenetics: a rejoinder on total evidence and consensus. *Systematic Biology* 50:881–891.
- LEVASSEUR, C. AND LAPOINTE, F.-J. 2002. A family of average consensus methods for weighted trees. In K. Jajuga, A. Sokolowski, and H.-H. Bock (eds), *Classification, Clustering and Data Analysis: Recent Advances and Applications*, pp. 365–369. Springer-Verlag, Berlin.
- LEVASSEUR, C. AND LAPOINTE, F.-J. 2003. Increasing phylogenetic accuracy with global congruence. In M. Janowitz, F.-J. Lapointe, F. R. McMorris, B. Mirkin, and F. S. Roberts (eds), *Bioconsensus*, pp. 221–230. American Mathematical Society, Providence, Rhode Island.
- LOSOS, J. B. 1994. An approach to the analysis of comparative data when a phylogeny is unavailable or incomplete. *Systematic Biology* 43:117–123.

- MAKARENKO, V. 2002. Comparison of four methods for inferring additive trees from incomplete dissimilarity matrices. In K. Jajuga, A. Sokolowski, and H.-H. Bock (eds), *Classification, Clustering and Data Analysis: Recent Advances and Applications*, pp. 371–378. Springer-Verlag, Berlin.
- MAKARENKO, V. AND LECLERC, B. 1999. The fitting of a tree metric according to a weighted least-squares criterion. *Journal of Classification* 16:3–26.
- MARGUSH, T. AND MCMORRIS, F. R. 1981. Consensus n -trees. *Bulletin of Mathematical Biology* 43:239–244.
- MCMORRIS, F. R. 1985. Axioms for consensus functions on undirected phylogenetic trees. *Mathematical Biosciences* 74:17–21.
- MIRKIN, B. AND ROBERTS, F. S. 1993. Consensus functions and patterns in molecular sequences. *Bulletin of Mathematical Biology* 55:695–713.
- RAMBAUT, A. AND GRASSLY, N. C. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer Applications in the Biosciences* 13:235–238.
- RAGAN, M. A. 1992. Phylogenetic inference based on matrix representations of trees. *Molecular Phylogenetics and Evolution* 1:53–58.
- SEMPLE, C. AND STEEL, M. 2000. A supertree method for rooted trees. *Discrete Applied Mathematics* 105:147–158.
- SMITH, T. J. 2001. Constructing ultrametric and additive trees based on the L-1 norm. *Journal of Classification* 18:185–207.
- SOKAL R. R. AND ROHLF, F. J. 1981. Taxonomic congruence in the Leptopodomorpha re-examined. *Systematic Zoology* 30:309–325.
- STINEBRICKNER, R. 1984. An extension of intersection methods from trees to dendograms. *Systematic Zoology* 33:381–386.
- STRIMMER, K. AND VON HAESELER, A. 1996. Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. *Molecular Biology and Evolution* 13:964–969.
- SWOFFORD, D. L. 2002. *PAUP**. *Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Version 4. Sinauer, Sunderland, Massachusetts.
- VICHI, M. 1993. Un algoritmo dei minimi quadrati per interpolare un insieme di classificazioni gerarchiche con una classificazione consenso. *Metron* 51:139–163.
- WILKINSON, M., LAPOINTE, F.-J., AND GOWER, D. J. 2003. Branch lengths and support. *Systematic Biology* 52:127–130.
- WILKINSON, M., THORLEY, J. L., LITTLEWOOD D. T. J., AND BRAY, R. A. 2001. Towards a phylogenetic supertree of Platyhelminthes? In D. T. J. Littlewood, and R. A. Bray (eds), *Interrelationships of the Platyhelminthes*, pp. 292–301. Taylor and Francis, London.

2. New supertree methods

Chapter 5

TANGLED TALES FROM MULTIPLE MARKERS

Reconciling conflict between phylogenies to build molecular supertrees

James A. Cotton and Roderic D. M. Page

Abstract: Supertree methods combine information from multiple phylogenies into a larger, composite phylogeny. When there is no disagreement between the source phylogenies, constructing the supertree is straightforward. But in the (nearly universal) presence of disagreement between source trees, supertree methods seek to either represent or resolve this conflict. Existing supertree methods that resolve conflict between source trees do so in an *ad hoc* way. Gene tree parsimony is a supertree method that can combine molecular phylogenies for overlapping taxon sets and interprets conflict between these phylogenies in a biologically meaningful way. We review the method and discuss the relationship between gene tree parsimony and other supertree methods. Finally, we suggest that a better understanding of the causes of conflict between source trees should lead to appropriate ways of resolving this conflict when constructing supertrees.

Keywords: gene duplication, gene tree parsimony, reconciled trees

1. Introduction

Combining information from different sources of phylogenetic evidence can be important for two different reasons: 1) to increase the scope of the phylogenetic results by including a greater range of terminal taxa, or 2) to improve the accuracy of the results by incorporating more data for these taxa. Supertree methods have been used to achieve both these aims by incorporating source trees constructed from a wide range of relevant data.

Where source trees are rooted and compatible, supertree construction is relatively trivial: efficient algorithms exist to decide whether or not a set of trees are compatible and to construct the parent trees that contain all these trees (Aho *et al.*, 1981; Steel, 1992; Semple, 2003). However, most practical applications of supertree methods involve source trees that are incompatible, and supertree workers have been less successful in designing algorithms to combine information from conflicting trees. Such algorithms can remove conflict by pruning leaves (e.g., in maximum agreement subtrees), represent the conflict through soft polytomies, resolve the conflict, or use some combination of these.

In fact, the only supertree method that has been at all used widely by biologists is matrix representation with parsimony (MRP; see Baum and Ragan, 2004), with an increasing number of supertrees constructed using this method appearing in the literature (e.g., Kennedy and Page, 2002; Pisani *et al.*, 2002; see Baum and Ragan, 2004). MRP uses additive binary coding to represent the hierarchical structure of a set of trees as a series of matrix elements — each node on the trees is represented by a column of the matrix, with missing data for those taxa not present on a particular source tree. This matrix is then analyzed using parsimony methods to construct a supertree or set of supertrees. Although MRP supertrees have played an important part in stimulating the field of supertree research and might be reasonably successful in reconstructing relationships (Bininda-Emonds and Sanderson, 2001), there has been an increasing literature on the biases of MRP methods, and several proposed modifications to the original method (e.g., Purvis, 1995; Ronquist, 1996; Bininda-Emonds and Bryant, 1998; Thorley, 2000). There are similar problems with other supertree algorithms, such as the MINCUTSUPERTREE method (Semple and Steel, 2000), which has several undesirable properties (Page, 2002). These problems have prompted a widening interest in other methods of supertree construction, such as shown in this volume and elsewhere (Page, 2002).

In an effort to classify the growing number of supertree methods available to systematists, at least two authors have characterized the supertree problem in a distance framework (Chen *et al.*, 2003; Thorley and Wilkinson, 2003). These authors suggest that the supertree problem can be seen as the problem of finding a tree (or set of trees) that is closest to a set of input trees under some measure of distance between trees. For example, as both sets of authors point out, MRP seeks to find the tree minimizing the number of steps required on the MRP matrix. Other distance measures are certainly possible, such as distances based on nearest-neighbour interchanges (NNIs; Waterman and Smith, 1978). Bearing this framework in mind, we note that all problems of identifying an optimal tree are likely to be NP-complete (Wareham, 1993), including the maximum-parsimony problem

used by MRP methods (Graham and Foulds, 1982). Thus, heuristic strategies are likely to be needed.

In this framework, we suggest a new distance measure for supertree inference, one based on the number of actual biological events that might have produced the differences observed between source trees. These events can be inferred using the co-phylogenetic method of reconciled trees. In this chapter, we introduce reconciled trees and their use to infer a species tree, or supertree, from several molecular source trees, a procedure which has become known as gene tree parsimony (GTP; Slowinski and Page, 1999). We include a brief empirical example of a GTP supertree. We then make a preliminary attempt to characterize the GTP method by describing some properties of the method, as has been attempted for other supertree methods. Lastly, we go beyond GTP itself to argue that understanding the causes of conflict between source trees should help us resolve that conflict appropriately, and to suggest that a model-based framework might enable systematic biologists both to understand the causes of conflict between trees and to construct accurate supertrees in the face of such conflict.

2. Tangled trees, or co-phylogeny

Evolutionary biologists have long been interested in the relationship between ecologically associated entities, particularly hosts and their parasites. One important question in host-parasite biology is the extent to which these organisms co-evolve, and, more specifically, the extent to which they co-diverge (i.e., the extent to which speciation events in one lineage are mirrored by speciation events in the other). This led to interest in comparing the phylogenetic trees of associated organisms, along with a parallel interest in relating the phylogenies of organisms to their biogeography (Page and Charleston, 1998). The initial solution to this problem was to use a binary coding of the dependant tree, similar to those used in MRP supertree methods. This matrix was then used either to reconstruct the host phylogeny, or to understand the pattern of evolution by optimizing the characters onto the second phylogeny (Brooks, 1981). Similar to the problems with the binary coding used in MRP, various fixes failed to alleviate the fundamental problem that the characters produced by this coding for a given tree are non-independent.

In studies of co-phylogeny, the solution has been to map the dependant phylogeny explicitly into the host phylogeny, postulating directly events that lead to the differences between the two phylogenies (see Figure 1). This insight led to Page's (1994) formalization of the earlier concept of reconciled trees (introduced by Goodman *et al.*, 1979). Constructing a reconciled tree

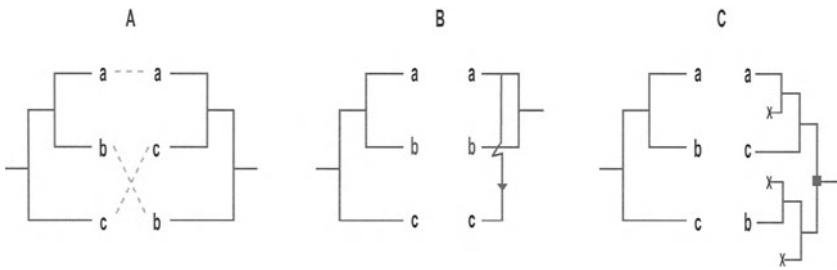


Figure 1. The incongruence between the species tree (A, left) and gene tree (A, right) in this example can be explained by postulating either a single lateral gene transfer from taxon a to taxon c (B) or a single gene duplication followed by three gene losses (C).

involves reconciling the differences between two trees by postulating certain co-phylogenetic events that introduced these differences. As shown in Figure 2, these events can be extinction of a lineage, independent speciation of a lineage, and horizontal transfer. Although co-phylogeny methods were developed in the context of biogeography and host-parasite evolution, similar events occur in the evolution of a gene lineage within a species (e.g., lateral gene transfer, gene duplications, and gene loss), so the same co-phylogeny mapping can also be used to study this system. Other evolutionary processes are also included under these co-phylogenetic events, with, for example, hybridization and some forms of recombination being indistinguishable from lateral gene transfer in this context.

The interest in supertree methods underlines the growing availability of phylogenies, and this increasing amount of data reflects both an increase in the taxonomic coverage of phylogenetic information (“width”) and in the amount of data available for particular organisms (“depth”). This increasing depth is particularly a result of the rise of genome-level sequencing efforts for an increasing number of organisms, and an important corollary of this work is the increasing realization that phylogenies for different genetic loci for the same species frequently disagree. This has in turn prompted the realization that a range of evolutionary events can cause the correct phylogeny for a gene to be different from the correct phylogeny for the species it is sampled from, a problem known as the gene tree-species tree problem (Doyle, 1992; Maddison, 1997). Reconciled trees are a natural solution to this problem (Page and Charleston, 1997a) — we can use the reconciled-tree algorithm to score a species tree for a particular gene tree in terms of the number of gene duplications, gene losses and other evolutionary events that have introduced differences between the two trees. The numbers of these events is a distance between the trees that has a natural, biological interpretation (Mirkin *et al.*, 1996).

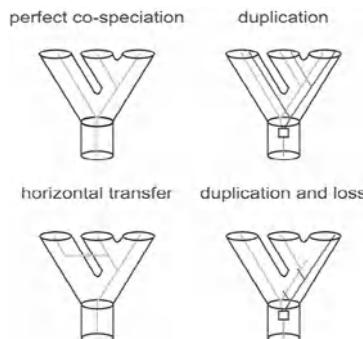


Figure 2. Some co-phylogenetic events, introducing differences between two associated phylogenies.

In principle, several different events can be scored in this way (Figure 2), including the number of deep coalescence events (Maddison, 1997). It should be noted that dealing with horizontal gene transfer correctly is complex and existing implementations of reconciled trees in this context exclude this possibility (Page, 1998). In particular, including horizontal transfer events makes tree reconciliation far more intensive computationally, and requires additional assumptions about the relative rates of gene duplication and loss and lateral gene transfer. Fortunately, solutions for co-phylogeny mapping incorporating horizontal transfer are available, and could be used in the context of GTP (Charleston, 1998; Ronquist and Nylin, 1990; Ronquist, 2003). Methods are also available for estimating optimal event costs for particular problems that suggest that reconciliation methods are robust to alternative weighting of different events (Ronquist, 2003). Even if just duplications and losses are included in the event set, different weightings of these two events are possible and will affect the result obtained (Ronquist, 2003). Fortunately, it seems that the duplication-and-loss optimal trees are a subset of the duplication-only optimal trees for a particular set of source trees (Page and Charleston, 1997b), so the consensus results with different weightings will differ only in degree of resolution. It is also often preferable to use the count of duplications alone (ignoring gene losses) as a distance function because gene losses are confounded with failure-to-sample in some kinds of study (e.g., due simply to the lack of a sequence in the sequence databases), and so do not represent a true biological cost (Cotton and Page, 2003). For the remainder of this chapter, we restrict ourselves to GTP using only duplication events or the sum of duplication and loss events, for which a software implementation is available (Page, 1998).

3. From reconciled trees to supertrees

When we have multiple gene trees, we can combine information from all these trees into a single tree by finding the species tree (or set of species trees) that minimizes the number of co-phylogenetic events required to reconcile the species tree with each source tree or minimizes some weighted sum of these events (assigning a cost to each event category). The resultant species tree can be on a larger taxon set than any of the source trees and is constructed using information from the topology of each source tree only. As such, it fits the definition of a conventional supertree. The set of GTP supertrees is thus the set of all supertrees that require a minimum number of the evolutionary events considered to explain the difference between the supertree and the set of source trees.

Finding an optimal species tree under either the duplication-only or duplication-and-loss score has been the focus of some attention by mathematicians and computational biologists. Linear-time algorithms exist for computing these scores for a particular pair of gene and species trees (Eulenstein, 1997; Zhang, 1997; Zmasek and Eddy, 2001), and although it is known (as expected) that finding the minimum-cost species tree is NP-complete (Ma *et al.*, 1998), there is a polynomial-time (fixed-parameter tractable) algorithm to find this tree where the maximum number of gene lineages extant at any point on the tree has an upper bound (Hallett and Lagergren, 2000).

If we restrict the source trees to be molecular trees, the duplication count (or duplication cost) is a biologically interpretable measure of the evolutionary difference between the source tree (or gene tree) and supertree (or species tree). If all the differences between source trees were a result of the evolutionary events included, then the GTP supertree would be expected to reconstruct the correct supertree accurately (at least as far as the methodological assumptions of parsimony hold). Unfortunately, little is understood about the causes of disagreement between molecular phylogenies. Clearly, some error will be a result of simple estimation error owing to the finite amount of data available from any single gene. The inadequacy of existing models will also lead to some error and so introduce conflict between phylogenies. Thus, it could be that little of the error between phylogenetic estimates from different molecular markers is because of the kinds of evolutionary events dealt with by GTP, and it is unclear how GTP will perform at resolving conflict from other (non-molecular) sources. It is, however, similarly unclear exactly how well other supertree methods perform in practice, although a start has been made on using simulation studies to address this for some methods (Bininda-Emonds and Sanderson, 2001, Burleigh *et al.*, 2004; Lapointe *et al.*, 2004). It is clearly an empirical

question how well any supertree method performs in practice, and there seems no reason to suspect that GTP will necessarily underperform compared with other methods when phylogenetic conflict is a result of estimation error or model inadequacy. More work is needed in comparing supertree methods in a range of situations before the strengths and weaknesses of different supertree methods will be understood.

One modification to standard supertree methods that has been shown to be highly effective in improving the accuracy of results (Ronquist, 1996; Bininda-Emonds and Sanderson, 2001; Salamin *et al.*, 2002) involves incorporating some measure of uncertainty into the input source trees (e.g., from a bootstrap profile of trees from non-parametric bootstrapping). An idea akin to this “weighted MRP” has also been mentioned in the reconciled-tree literature, where it seems particularly apposite. If reconciled-tree methods rely on identifying evolutionary events that lead to incongruence between trees, it is clearly crucial to incorporate some idea of the uncertainty in tree estimates if these events are to be “real” rather than owing to this uncertainty (Page, 2000; Page and Cotton, 2000; Ronquist, 2003). Using a bootstrap profile of trees for each gene has been shown to improve the species tree estimate in at least one empirical study (Cotton and Page, 2002), and also provides analogous bootstrap support values for the species tree or supertree itself. Several other methods for incorporating uncertainty in source tree estimates into reconciled tree analyses have also been proposed (Page, 2000; Page and Cotton, 2000).

4. An empirical example: a small supertree of *Drosophila*

Several empirical examples of using reconciled-tree methods to infer phylogenies exist in the literature (Slowinski *et al.*, 1997; Page, 2000; Cotton and Page, 2002; Martin and Burg, 2002), but we present here a novel empirical example of a small-scale supertree of *Drosophila* and some related genera based on five nuclear genes (Figure 3). The source trees were relabeled with the species names and a standard MRP matrix was built using the program Supertree (available at <http://darwin.zoology.gla.ac.uk/~rpage/supertree/>). The MRP matrix was analyzed using PAUP* v4b10 (Swofford, 2002) using standard parsimony. The GTP analysis was performed using GeneTree (Page, 1998). For both analyses, a large number of equally optimal trees were found, so five separate searches were performed, with each one swapping on a maximum of 50 000 (for MRP) or 15 000 (for GTP) trees. Consensus trees for each of the five searches were very similar,

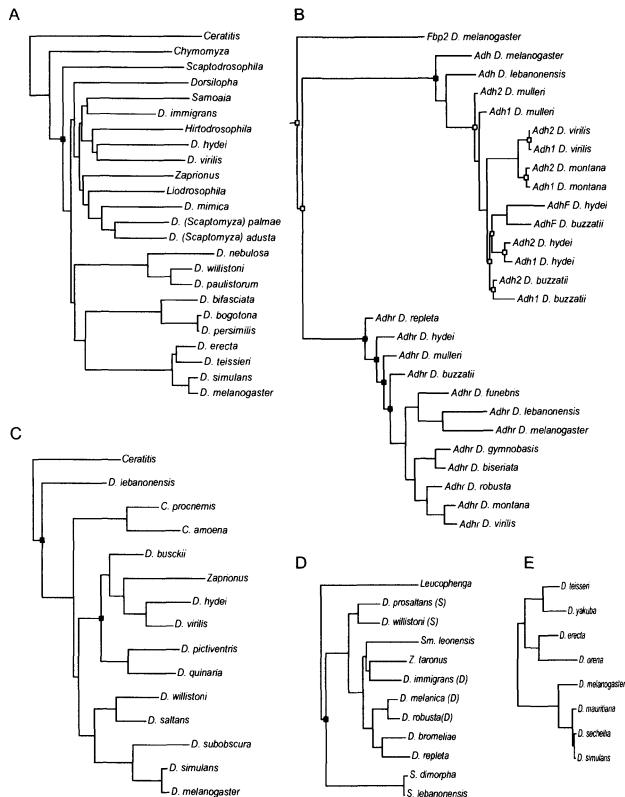


Figure 3. The five gene trees used in building the *Drosophila* supertree presented here: A) *Dopa decarboxylase* (Tatarenkov *et al.*, 1999), B) *Alcohol dehydrogenase* and the *Alcohol dehydrogenase-related gene* (Betrán and Ashburner, 2000), C) *Cu-Zn superoxide dismutase* (Kwiatowski *et al.*, 1994), D) 28S rRNA (Russo *et al.*, 1995), and E) the regulatory gene *roughex* (Avedisov *et al.*, 2001). Boxes show positions of gene duplications implied by the supertrees. Open boxes are duplications necessitated by the multiple copies of *Alcohol dehydrogenase* genes, whereas closed boxes are those duplications inferred from conflict between the gene tree and the supertree. All the duplications, except that for 28S rRNA, are implied by every supertree.

suggesting that the five searches had each sampled successfully from across the large island of trees. Trees of cost 97 parsimony steps were found under MRP, and 63 duplications and losses under GTP. The set of GTP supertrees thus included all the trees reconciled with the five source trees using all combinations of 63 duplications and losses (in fact, all the supertrees found required either 17 duplications and 46 losses, or 18 duplications and 45 losses). Although 18 duplications sounds like a lot, nine duplications are required by multiple gene copies being present on the

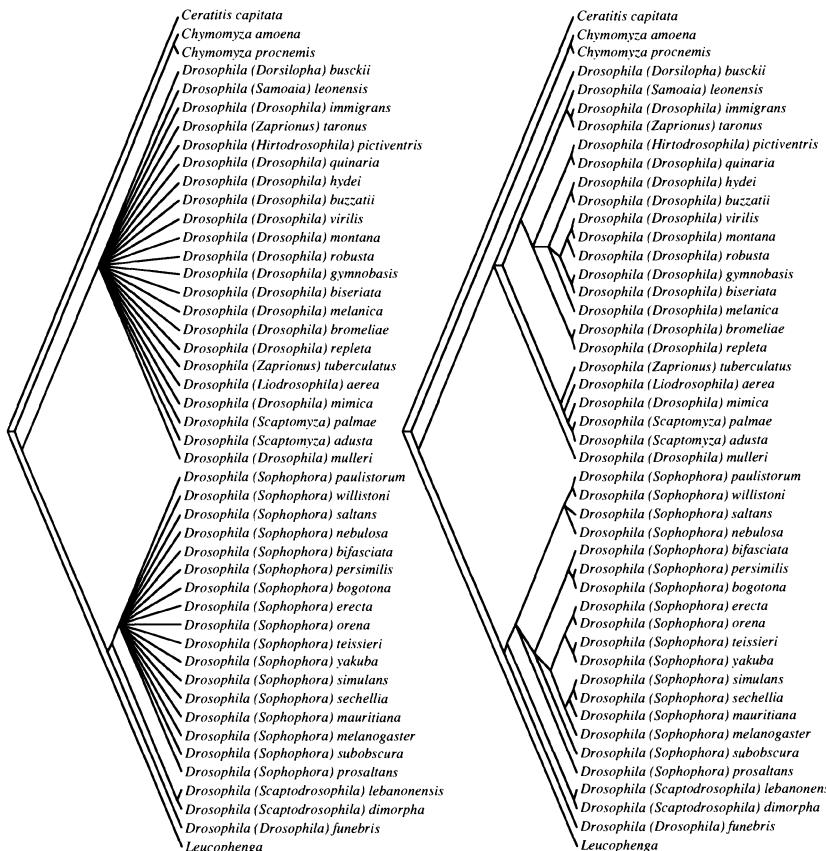


Figure 4. The strict component (left) and Adams (right) consensus of the GTP supertrees from the *Drosophila* gene trees under the duplication-and-loss criterion.

Alcohol dehydrogenase gene tree (Figure 3). Thus, only nine duplications, at most, are a result of incongruence between the source trees and supertrees.

Given that they are derived from the same data, it is reassuring that both the GTP and MRP analyses are similar (Figures 4 and 5, respectively). Both analyses support the monophyly of the subgenus *Sophophora*, and indeed show exactly the same relationships within *Sophophora*. It appears that GTP is more conservative than MRP, in that the results it produces are largely compatible with those from MRP, but somewhat less resolved (although this is not the case for every clade). Both GTP and MRP find the other subgenera of *Drosophila* included to be paraphyletic or polyphyletic, but there are some differences between the two sets of trees. One instructive difference is in the way *D. melanica*, *D. robusta*, *D. biseriata* and *D. gymnobasis* cluster

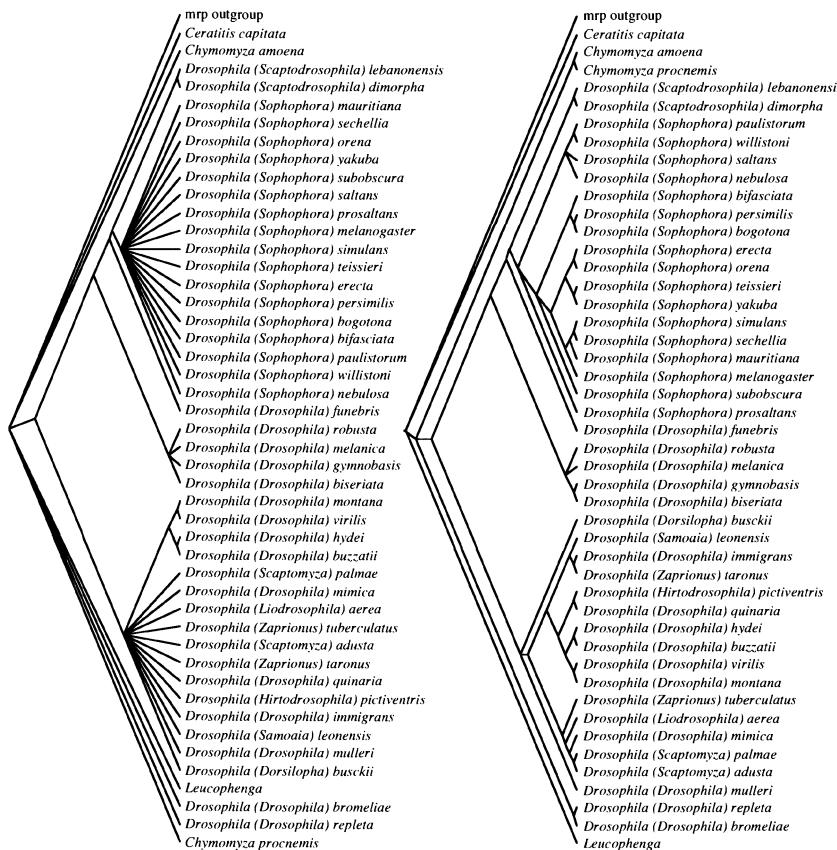


Figure 5. The strict component (left) and Adams (right) consensus of the standard MRP supertrees.

with respect to one another. In the MRP results, these species are placed strikingly away from the other members of subgenus *Drosophila* as a sister clade to *D. funebris* and the subgenera *Sophophora* and *Scaptodrosophila*. This is in contrast to the GTP tree, where these species are embedded within the paraphyletic assemblage around subgenus *Drosophila* to which they all belong. The MRP results seem surprising given the source trees: the four species in question appear only on trees in Figures 3B and D, where they group with other members of the subgenus *Drosophila*. This odd placement is probably partly a result of the different treatment of *D. bromeliae* and *D. repleta*, the other principal difference between the two sets of trees. These two species are the sister taxa to *D. melanica* and *D. robusta* on the 28S tree (Figure 3D). The unresolved position of *D. bromeliae* and *D. repleta* on the MRP tree is understandable given that they are placed (with three other

members of the clade containing subgenus *Drosophila* and others), between subgenera *Sophophora* and *Scaptodrosophila* on the 28S tree. However, given the support for the three other taxa as being related to members of the subgenus *Drosophila* on two other source trees (Figures 3A, C), the more-resolved position on the GTP trees seems at least as reasonable.

Investigators can examine incongruence in the GTP supertree in terms of duplications and losses in specific genes. This can both help assess whether incongruence is restricted to a single gene (i.e., because it contains the vast majority of duplications and losses) and help to understand the general pattern of genetic evolution for this group. Furthermore, the hypothesized duplications and losses might be testable using other evidence: for example, do the suggested paralogues have different functions, occur in different parts of the genome, or have different genetic architectures? Another approach might be to use the GTP supertree to inform a search for additional gene copies. For example, the proposed duplication in *Dopa decarboxylase* could be confirmed by finding an additional copy of the gene in *Scaptodrosophila*, although it would be wise to examine the strength of support for a particular duplication before expending much laboratory effort on such a search!

5. Properties of GTP as a supertree method

Progress has been made recently in thinking about desirable properties of supertree methods (see Wilkinson *et al.*, 2004). These properties are characteristics that would seem to be desirable in all supertree methods, and which seem likely to correlate with the accuracy of the results of a method. Comparatively little has been done to characterize supertree methods formally in terms of these properties or more formal axioms. In particular, it might be of interest to see how GTP resolves conflict between source trees when compared with those variants of MRP that are already characterized in terms of some of these properties. The properties named in italics below are used in the sense of Wilkinson *et al.* (2004). Aside from the three properties discussed below, GTP methods are *assessable*, *weightable*, *plenary*, show *order invariance*, and seem to be *Pareto* on components. They do not show *generality* or *uniqueness*, and are not particularly *speedy* compared with polynomial-time methods. Their behaviour in terms of being *co-Pareto* and *independent of irrelevant alternatives* is unclear.

5.1 GTP displays unique subtrees correctly

Here, I define a unique subtree as one that appears in a single source tree, where no other source tree contains any of the taxa of the subtree. GTP

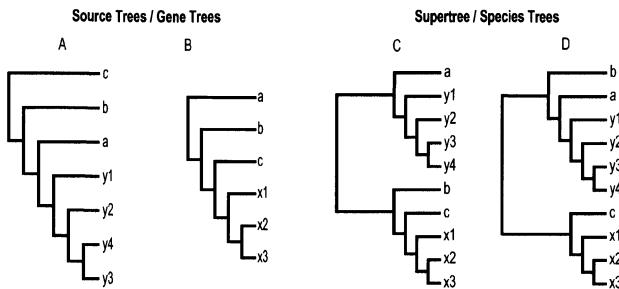


Figure 6. Trees C and D are the two supertrees for source trees A and B under both the duplication-only and duplication-and-loss costs. Trees C and D are also the standard and Purvis coding MRP supertrees for trees A and B (source trees taken from Page, 2002).

appears to include unique subtrees in the supertree or species tree correctly, a property shared by MRP methods, but not by the original formulation of MINCUTSUPERTREE (Page, 2002). Using Page's example (Figure 6), we see that GTP reconstructs these groupings correctly under both duplication-only and duplication-and-loss criteria. Both GTP and MRP perform better than the modified MinCut method in placing taxon a correctly as sister-group to the clade (x_1, \dots, x_3) and taxon c as sister-group to the clade (y_1, \dots, y_4), rather than collapsing these relationships to a polytomy (Page, 2002). Clearly, reconstructing clades that are unique to a single tree is a desirable property for all supertree methods. This property is a special case of property P7 of Steel *et al.* (2000), which they showed no rooted supertree method that produces a single output tree can possess.

5.2 GTP is not *sizeless*

It has been noted that the original coding for MRP matrices produces supertrees biased towards including those relationships on larger source trees because of redundant information in the matrix (Purvis, 1995). Purvis showed that some matrix entries are redundant in the sense of not being needed to reconstruct the original source trees, but this information might not be redundant in a different sense (see Ronquist, 1996). We use Purvis's example to show that GTP also suffers from this bias when the duplication-and-loss criterion is used, but not under the duplication-only criterion. The two gene trees shown in Figures 7A and B support just a single species tree under the duplication-and-loss criterion, that of Figure 7C. In this tree, taxon d occurs in the position supported by tree A, the larger of the two source trees, thereby ignoring effectively the conflicting signal from the very different position of this taxon in the smaller tree B. Under the duplication-

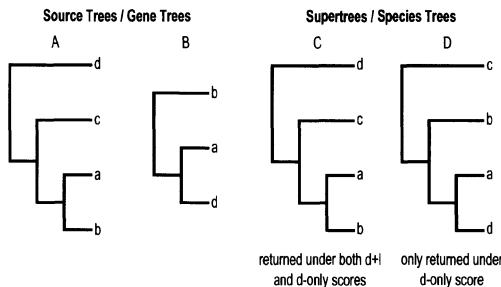


Figure 7. Trees C and D are the two supertrees for source trees A and B under the duplication-only cost. Tree C is the unique supertree under the duplication-and-loss cost. Tree C is the unique supertree under standard MRP, while both C and D are Purvis-coding MRP supertrees (source trees taken from Purvis, 1995).

only criterion, an additional species tree (Figure 7D) has an equal cost and shows taxon d in the position suggested by the smaller input tree.

The reason for this bias under the duplication-and-loss criterion is clear: duplications inferred on larger gene trees will tend to infer more gene losses than those on smaller trees. Under this criterion, the species tree will thus be selected to minimize gene duplications on larger gene trees more than on smaller ones, and so will tend to reflect relationships in larger gene trees. This source of bias disappears under the duplication-only criterion.

5.3 GTP is not *positionless*

Several suggested variants of MRP appear to suffer from a bias towards placing species in the most crownward position displayed by the input trees. This bias was first noticed by Ronquist (1996) as being a problem with Purvis's (1995) suggested modification to the original MRP encoding. Figure 8 shows two source trees, A and B. Under both the duplication-and-loss and duplication-only criteria, there is only a single optimal species tree (Figure 8C). This tree places taxon e in the more crownward position, as suggested by source tree B, overruling the conflicting position suggested by source tree A. Thus, it seems that GTP also shows a bias towards placing taxa in the more crownward position.

6. A probabilistic view of the supertree problem

We can view the supertree problem usefully in a probabilistic setting, a view that makes several themes of this paper particularly clear. This is a fairly natural extension of the distance-based view expressed earlier. Instead of

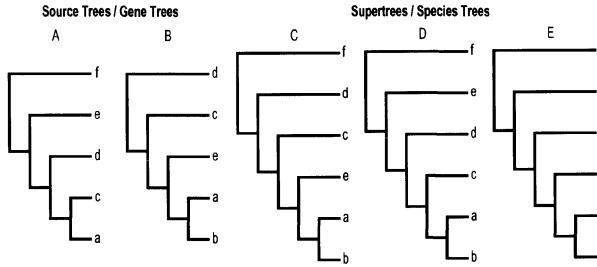


Figure 8. Tree C is the unique GTP supertree for source trees A and B under both duplication-only and duplication-and-loss costs. Tree C is also the unique MRP supertree under Purvis coding, while the all three trees C to E are MRP supertrees under standard coding. Adapted from Thorley (2000).

seeking the closest tree to a set of source trees, we can look for the maximum likelihood or most probable supertree for this set. To do this, we need a likelihood function for the supertree that is proportional to the probability that the source trees come from the supertree. There are several different ways we could frame this likelihood function based on how similar the source trees are to the subtrees of the proposed supertree induced by their leaf sets. For example, if we assume every NNI needed to move from the induced subtree to the source tree is equally likely, it is relatively trivial to construct this function using a binomial distribution. To do this, we need only calculate the NNI distance between source tree and its induced subtree in the supertree, and the maximum possible distance between the trees under this operation. The product of these probabilities across all sources trees would then be the likelihood of this supertree under this simple NNI-binomial model. The model has only a single parameter that must be estimated from the data: q , the probability of an NNI difference between a source tree and the supertree. The likelihood of a supertree T_s from a set of n subtrees $T_1 \dots T_n$, where the NNI distance between the source tree T_i and the subtree induced on T_s by the leaves of T_i is d_{T_i, T_s} and the maximum NNI distance between two trees of this size is ΔG , is given by

$$(1) \quad L(T_s | T_1, T_2, \dots, T_n) \propto \prod_{i=1}^n p(T_i | T_s),$$

where the probability of each source tree is simply

$$(2) \quad p(T_i | T_s, q) = \binom{\Delta G}{d_{T_i, T_s}} q^{d_{T_i, T_s}} (1-q)^{\Delta G - d_{T_i, T_s}}.$$

Constructing this likelihood function allows us to find a maximum likelihood supertree under this model using standard heuristic methods. It would also be easy to estimate the supertree in a Bayesian framework using Markov chain Monte Carlo. To do this, we need to propose a prior probability distribution on the supertree and place a prior on the NNI probability parameter of the model. A Bayesian method would let us construct a credible interval of trees within which the true supertree lies with high probability. Sampling from this posterior probability distribution of supertrees will also allow the use of correct probability distributions for trees, improving the accuracy of the various evolutionary studies in which supertrees have been used (Huelsenbeck *et al.*, 2000b; also Ronquist *et al.*, 2004). It should be noted that the above discussion assumes that source trees are known without error. If character data are available for all the subtrees, an obvious approach would be to calculate the probabilities of these trees using a model of sequence evolution, providing a natural way to incorporate uncertainty in the source tree estimates.

More importantly, formulating the supertree problem in this way shows that a wide range of likelihood functions relating a subtree to the supertree could be used to build supertrees. We emphasize that models such as the NNI-binomial model are likely to be gross simplifications and inadequate for most estimation purposes, so more complex models (such as that of Ronquist *et al.*, 2004) will be needed. More interestingly, probabilistic models of gene duplication and gene loss have been developed recently (Arvestad *et al.*, 2003) that could be extended to the supertree setting. Even horizontal transfer events can be incorporated (Huelsenbeck *et al.*, 2000a), although this is more difficult to model mathematically (Charleston and Robertson, 2002). It seems probable that simplifying assumptions akin to those of single base substitutions in DNA sequence phylogeny models will be needed for supertree models. Perhaps the greatest advantage of both likelihood and Bayesian methods is that both provide a natural framework for comparing models, and so permit rational choice between different methods. As discussed earlier, relatively little is known about how different methods perform on real data, and it could be in this probabilistic framework that competing methods, and their different assumptions, can be compared the most rigorously. The simplistic model presented here might be a useful null model against which more realistic models can be tested.

7. Conclusion

We are clearly at an early stage in the development of supertree methods: many methods are being proposed, but little is known about their relative

merits. While most supertree methods treat conflict between source trees in an *ad hoc* way, it is possible to treat at least some causes of incompatibility in a biologically realistic way. We hope that this chapter will encourage biologists to think more about how incongruence between trees can be investigated, and about the possible causes of this incongruence beyond simple estimation error. It is clearly an empirical question how different supertree methods will perform on real data, and it is probable that different methods will be preferable for different data, reflecting the different causes of conflict in them. For example, reconciled-tree methods might be the most appropriate if all conflict between source trees is caused by gene duplication and gene loss (probably a rather unlikely scenario), whereas matrix representation with flipping (Burleigh *et al.*, 2004) might perform best where conflict results in randomly distributed errors on some binary matrix representation of the source trees. Much more work is clearly needed to understand both the causes and consequences of conflict between phylogenies from different data.

Acknowledgements

We thank Olaf Bininda-Emonds for inviting us to write this chapter and also Olaf Bininda-Emonds, Fredrik Ronquist, Gordon Burleigh, and Mark Wilkinson for comments on the manuscript. This work was performed while JAC was supported by a NERC studentship while at the Division of Environmental and Evolutionary Biology at University of Glasgow, and by BBSRC grant 40/G18385.

References

- AHO, A. V., SAGIV, Y., SZYMANSKI, T. G., AND ULLMAN, J. 1981. Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. *SIAM Journal of Computing* 10:405–421.
- ARVESTAD, L., BERGLUND, A.-C., LAGERGREN, J., AND SENNBLAD, B. 2003. Bayesian gene/species tree reconciliation and orthology analysis using MCMC. *Bioinformatics* 19:i7–i15.
- AVEDISOV, S. N., ROGOZIN, I. B., KOONIN, E. V., AND THOMAS, B. J. 2001. Rapid evolution of a cyclin A inhibitor gene, *roughex*, in *Drosophila*. *Molecular Biology and Evolution* 18:2110–2118.
- BAUM, B. R. AND RAGAN, M. A. 2004. The MRP method. In O. R. P. Bininda-Emonds (ed). *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 17–34. Kluwer Academic, Dordrecht, the Netherlands.
- BETRÁN, E. AND ASHBURNER, M. 2000. Duplication, dicistronic transcription, and subsequent evolution of the Alcohol dehydrogenase and Alcohol dehydrogenase-related genes in *Drosophila*. *Molecular Biology and Evolution* 17:1344–1352.

- BININDA-EMONDS, O. R. P. AND BRYANT, H. N. 1998. Properties of matrix representation with parsimony analyses. *Systematic Biology* 47:497–508.
- BININDA-EMONDS, O. R. P. AND SANDERSON, M. J. 2001. Assessment of the accuracy of matrix representation with parsimony analysis supertree construction. *Systematic Biology* 50:565–579.
- BROOKS, D. R. 1981. Hennig's parasitological method: a proposed solution. *Systematic Zoology* 30:229–249.
- BURLEIGH, J. G., EULENSTEIN, O., FERNÁNDEZ-BACA, D., AND SANDERSON, M. J. 2004. MRF supertrees. In O. R. P. Bininda-Emonds (ed). *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 65–85. Kluwer Academic, Dordrecht, the Netherlands.
- CHARLESTON, M. A. 1998. Jungles: a new solution to the host/parasite phylogeny reconciliation problem. *Mathematical Biosciences* 149:191–223.
- CHARLESTON, M. A. AND ROBERTSON, D. L. 2002. Preferential host switching by primate lentiviruses can account for phylogenetic similarity with the primate phylogeny. *Systematic Biology* 51:528–535.
- CHEN, D., DIAO, L., EULENSTEIN, O., FERNÁNDEZ-BACA, D., AND SANDERSON, M. J. 2003. Flipping: a supertree construction method. In M. Janowitz, F.-J. Lapointe, F. R. McMorris, B. Mirkin, and F. S. Roberts (eds), *Bioconsensus*, pp. 135–160. American Mathematical Society, Providence, Rhode Island.
- COTTON, J. A. AND PAGE, R. D. M. 2002. Going nuclear: vertebrate phylogeny and gene family evolution reconciled. *Proceedings of the Royal Society of London B* 269:1555–1561.
- COTTON, J. A. AND PAGE, R. D. M. 2003. Gene tree parsimony vs. uninode coding for phylogenetic reconstruction. *Molecular Phylogenetics and Evolution* 29:298–308.
- DOYLE, J. J. 1992. Gene trees and species trees: molecular systematics as one-character taxonomy. *Systematic Botany* 17:144–163.
- EULENSTEIN, O. 1997. A linear time algorithm for tree mapping. *Arbeitspapiere der GMD*, No. 1046.
- FELSENSTEIN, J. 1981. Evolutionary trees from DNA-sequences – a maximum likelihood approach. *Journal of Molecular Evolution* 17:368–376.
- GOODMAN, M., CZELUSNIAK, J., MOORE, G. W., ROMERO-HERRERA, A. E., AND MATSUDA, G. 1979. Fitting the gene lineage into its species lineage: a parsimony strategy illustrated by cladograms constructed from globin sequences. *Systematic Zoology* 28:132–168.
- GRAHAM, R. L. AND FOULDS, L. R. 1982. Unlikelihood that minimal phylogenies for a realistic biological study can be constructed in reasonable computation time. *Mathematical Biosciences* 60:133–142.
- HALLETT, M. T. AND LAGERGREN, J. 2000. New algorithms for the duplication-loss problem. In R. Shamir, S. Miyano, S. Istrail, P. Pevzner, and M. Waterman (eds), *RECOMB '00, Proceedings of the Fourth Annual International Conference on Computational Molecular Biology*, pp. 138–146. Association for Computing Machinery.
- HUELSENBECK, J. P., RANNALA, B., AND LARGET, B. 2000a. A Bayesian framework for the analysis of cospeciation. *Evolution* 54:352–364.
- HUELSENBECK, J. P., RANNALA, B., AND MASLY, J. P. 2000b. Accommodating phylogenetic uncertainty in evolutionary studies. *Science* 288:2349–2350.
- KENNEDY, M. AND PAGE, R. D. M. 2002. Seabird supertrees: combining partial estimates of procellariiform phylogeny. *The Auk* 119:88–108.

- KWIATOWSKI, J., SKARECKY, D., BAILEY, K., AND AYALA, F. J. 1994. Phylogeny of *Drosophila* and related genera inferred from the nucleotide sequence of the cu,zn sod gene. *Journal of Molecular Evolution* 38:443–454.
- LAPOINTE, F.-J. AND LEVASSEUR, C. 2004. Everything you always wanted to know about the average consensus, and more. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 87–105. Kluwer Academic, Dordrecht, the Netherlands.
- MA, B., LI, M., AND ZHANG, L. 1998. On reconstructing species trees from gene trees in term of duplications and losses. In S. Istrail, P. A. Pevzner, and M. S. Waterman (eds), *Proceedings of the Second Annual International Conference on Computational Biology (RECOMB 98)*, pp. 182–191. ACM, New York.
- MADDISON, W. P. 1997. Gene trees in species trees. *Systematic Biology* 46:523–536.
- MARTIN, A. P. AND BURG, T. M. 2002. Perils of paralogy: using hsp70 genes for inferring organismal phylogenies. *Systematic Biology* 51:570–587.
- MIRKIN, B., MUCHNIK, I., AND SMITH, T. F. 1996. A biologically consistent model for comparing molecular phylogenies. *Journal of Computational Biology* 2:493–507.
- PAGE, R. D. M. 1994. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Systematic Biology* 43:58–77.
- PAGE, R. D. M. 1998. GeneTree: comparing gene and species trees using reconciled trees. *Bioinformatics* 14:819–820.
- PAGE, R. D. M. 2000. Extracting species trees from complex gene trees: reconciled trees and vertebrate phylogeny. *Molecular Phylogenetics and Evolution* 14:89–106.
- PAGE, R. D. M. 2002. Modified mincut supertrees. In R. Guigó and D. Gusfield (eds), *Algorithms in Bioinformatics, Second International Workshop, WABI 2002, Rome, Italy, September 17–21, 2002, Proceedings*, pp. 537–552. Springer, Berlin.
- PAGE, R. D. M. AND CHARLESTON, M. A. 1997a. From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Molecular Phylogenetics and Evolution* 7:231–240.
- PAGE, R. D. M. AND CHARLESTON, M. A. 1997b. Reconciled trees and incongruent gene and species trees. In B. Mirkin, F. McMorris, F. Roberts, and A. Rzhetsky (eds), *Mathematical Hierarchies in Biology*, pp. 57–70. American Mathematical Society, Providence, Rhode Island.
- PAGE, R. D. M. AND CHARLESTON, M. A. 1998. Trees within trees: phylogeny and historical associations. *Trends in Ecology and Evolution* 13:356–359.
- PAGE, R. D. M. AND COTTON, J. A. 2000. GeneTree: a tool for exploring gene family evolution. In D. Sankoff and J. H. Nadeau (eds), *Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and the Evolution of Gene Families*, pp. 525–536. Kluwer Academic Publishers, Dordrecht, The Netherlands.
- PISANI, D., YATES, A. M., LANGER, M. C., AND BENTON, M. J. 2002. A genus-level supertree of the Dinosauria. *Proceedings of the Royal Society of London B* 269:915–921.
- PURVIS, A. 1995. A modification to Baum and Ragan's method for combining phylogenetic trees. *Systematic Biology* 44:251–255.
- RONQUIST, F. 1996. Matrix representation of trees, redundancy, and weighting. *Systematic Biology* 45:247–253.
- RONQUIST, F. 2003. Parsimony analysis of coevolving species associations. In R. D. M. Page (ed.), *Tangled Trees: Phylogeny, Cospeciation and Coevolution*, pp. 22–64. University of Chicago Press, Chicago.

- RONQUIST, F., HUELSENBECK, J. P., AND BRITTON, T. 2004. Bayesian supertrees. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 193–224. Kluwer Academic, Dordrecht, the Netherlands.
- RONQUIST, F. AND NYLIN, S. 1990. Process and pattern in the evolution of species associations. *Systematic Zoology* 39:323–344.
- RUSSO, C. A. M., TAKEZAKI, N., AND NEI, M. 1995. Molecular phylogeny and divergence times of drosophilid species. *Molecular Biology and Evolution* 12:391–404.
- SALAMIN, N., HODKINSON, T. R., AND SAVOLAINEN, V. 2002. Building supertrees: an empirical assessment using the grass family (Poaceae). *Systematic Biology* 51:136–150.
- SEMPLE, C. 2003. Reconstructing minimal rooted trees. *Discrete Applied Mathematics* 127:489–503.
- SEMPLE, C. AND STEEL, M. 2000. A supertree method for rooted trees. *Discrete Applied Mathematics* 105:147–158.
- SLOWINSKI, J. AND PAGE, R. D. M. 1999. How should species phylogenies be inferred from sequence data? *Systematic Biology* 48:814–825.
- SLOWINSKI, J. B., KNIGHT, A., AND ROONEY, A. P. 1997. Inferring species trees from gene trees: a phylogenetic analysis of the Elapidae (Serpentes) based on the amino acid sequences of venom proteins. *Molecular Phylogenetics and Evolution* 8:349–362.
- STEEL, M. 1992. The complexity of reconstructing trees from qualitative characters and subtrees. *Journal of Classification* 9:91–116.
- STEEL, M., DRESS, A. W. M., AND BÖCKER, S. 2000. Simple but fundamental limitations on supertree and consensus tree methods. *Systematic Biology* 49:363–368.
- SWOFFORD, D. L. 2002. *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Version 4. Sinauer, Sunderland, Massachusetts.
- TATARENKOV, A., KWiatowski, J., SKARECKY, D., BARRIO, E., AND AYALA, F. J. 1999. On the evolution of *Dopa decarboxylase* (Ddc) and *Drosophila* systematics. *Journal of Molecular Evolution* 48:445–462.
- THORLEY, J. L. 2000. *Cladistic Information, Leaf Stability and Supertree Construction*. Ph.D. dissertation, University of Bristol.
- THORLEY, J. L. AND WILKINSON, M. 2003. A view of supertree methods. In M. Janowitz, F.-J. Lapointe, F. R. McMorris, B. Mirkin, and F. S. Roberts (eds), *Bioconsensus*, pp. 185–193. American Mathematical Society, Providence, Rhode Island.
- WAREHAM, H. T. 1993. On the computational complexity of inferring evolutionary trees. Technical Report 9301, Department of Computer Science, Memorial University of Newfoundland.
- WATERMAN, M. S. AND SMITH, T. F. 1978. On the similarity of dendograms. *Journal of Theoretical Biology* 73:789–800.
- WILKINSON, M., THORLEY, J. L., PISANI, D., LAPOINTE, F.J., AND MCINERNEY, J. O. 2004. Some desiderata for liberal supertrees. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 227–246. Kluwer Academic, Dordrecht, the Netherlands.
- ZHANG, L. 1997. On a Mirkin-Muchnik-Smith conjecture for comparing molecular phylogenies. *Journal of Computational Biology* 4:177–187.
- ZMASEK, C. M. AND EDDY, S. R. 2001. A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics* 17:821–828.

Chapter 6

SUPERTREE METHODS FOR ANCESTRAL DIVERGENCE DATES AND OTHER APPLICATIONS

David Bryant, Charles Semple, and Mike Steel

Abstract: There are many ways to combine rooted phylogenetic trees with overlapping leaf sets into a single “supertree”. The most widely used method is MRP (matrix representation with parsimony analysis), but other direct methods have been developed recently. However, all these methods utilize typically only the discrete topology of the input trees and ignore other information that might be available. Based, for example, on fossil data or molecular dating techniques, this information includes whether one particular divergence event occurred earlier or later than another, and actual time estimates for divergence events. The ability to include such information in supertree construction could allow for more accurate dating of certain species divergences. This is a topical problem in recent biological literature. In this chapter, we describe a way to incorporate divergence time information in a fast and exact supertree algorithm that extends the classic BUILD algorithm. The approach is somewhat flexible in that it allows any combination of relative and/or absolute divergence times. In addition to this extension, the last section of this chapter consists of applications of BUILD to problems in phylogenetics that are, in general, computationally challenging.

Keywords: BUILD; divergence dates; rooted phylogenetic tree; supertree

1. Introduction

We will follow mostly the notation of Semple and Steel (2003) and assume that the reader is familiar with the basic concepts of phylogenetic trees.

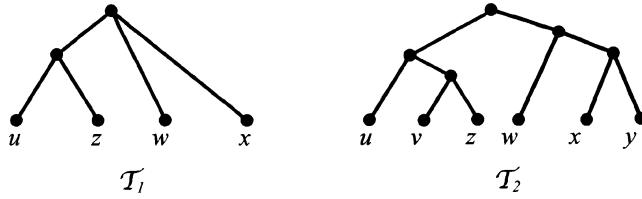


Figure 1. T_2 displays T_1 .

Let \mathcal{T} be a rooted phylogenetic X -tree. Thus, X refers to the leaf set of \mathcal{T} , here denoted $\mathcal{L}(\mathcal{T})$, which represents typically the set of extant species classified by \mathcal{T} . If $X' \subseteq X$, then the *restriction of \mathcal{T} to X'* , denoted $\mathcal{T}|X'$, is the rooted phylogenetic X' -tree that is obtained from the minimal rooted subtree of \mathcal{T} containing X' by suppressing all non-root vertices of degree two.

Let T_1 and T_2 be two rooted phylogenetic trees with $\mathcal{L}(T_1) \subseteq \mathcal{L}(T_2)$. We say that T_2 *displays* T_1 if $T_2| \mathcal{L}(T_1)$ is a refinement of T_1 . Informally, T_2 displays T_1 if, up to polytomies, all the ancestral relationships of T_1 are preserved in T_2 . This notion is illustrated in Figure 1.

For a collection \mathcal{R} of rooted phylogenetic trees, we denote the set $\cup_{\mathcal{T} \in \mathcal{R}} \mathcal{L}(\mathcal{T})$ by $\mathcal{L}(\mathcal{R})$. Extending the notion of display to \mathcal{R} , we say that a rooted phylogenetic tree \mathcal{T} with $\mathcal{L}(\mathcal{R}) \subseteq \mathcal{L}(\mathcal{T})$ *displays* \mathcal{R} if every member of \mathcal{R} is displayed by \mathcal{T} . Via an algorithm called BUILD, Aho *et al.* (1981) showed that there is a polynomial-time algorithm to determine if a rooted phylogenetic tree exists that displays \mathcal{R} and, if so, to construct such a rooted phylogenetic tree.

Informally, BUILD constructs clusters (subsets of $\mathcal{L}(\mathcal{R})$) that are broken down successively into disjoint subclusters according to the hierarchical relationships described by the trees in \mathcal{R} . If this process continues until all singleton clusters (i.e., clusters consisting of just a single species) are obtained, the rooted tree that corresponds to this resulting set of clusters is the tree returned by BUILD that displays \mathcal{R} . However, if the process stops before all singleton clusters are obtained, then there is no rooted phylogenetic tree that displays \mathcal{R} .

We remark here that BUILD was applied originally to relational databases and its application to phylogenetics appeared somewhat later (e.g., Steel, 1992; Constantinescu, 1995; Ng and Wormald, 1996; Semple, 2003). This fact might explain why BUILD is not that well known in this field.

A natural extension of BUILD is to include as input a collection of constraints representing the order in which the divergence events of certain different pairs of species occurred. To make this precise, we need several further definitions. Let \mathcal{T} be a phylogenetic tree. For all $v_1, v_2 \in V(\mathcal{T})$, we write $v_1 \leq_{\mathcal{T}} v_2$ if v_2 is a descendant of v_1 . The relation $\leq_{\mathcal{T}}$ induces a partial

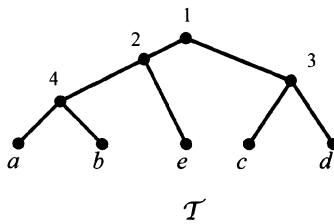


Figure 2. A ranked phylogenetic tree.

order on the vertices of \mathcal{T} . For a subset A of $\mathcal{L}(\mathcal{T})$, the unique vertex of \mathcal{T} that is the greatest lower bound of A under $\leq_{\mathcal{T}}$ is referred to as the *most recent common ancestor* of A in \mathcal{T} . We denote this vertex as $\text{mrca}_{\mathcal{T}}(A)$. For simplicity, if $A = \{a, b\}$, we denote $\text{mrca}_{\mathcal{T}}(\{a, b\})$ by $\text{mrca}_{\mathcal{T}}(a, b)$.

Let \mathcal{T} be a rooted phylogenetic tree. A *rank function* for \mathcal{T} is a function r from the set V^o of interior vertices of \mathcal{T} into the set of positive integers such that, for all $v_1, v_2 \in V^o$, $r(v_1) < r(v_2)$ if v_2 is a proper descendant of v_1 . The pair (\mathcal{T}, r) is a *ranked phylogenetic tree*. Again for simplicity, we denote $r(\text{mrca}_{\mathcal{T}}(A))$ by $r(A)$ for all subsets $A \in \mathcal{L}(\mathcal{T})$. In this chapter, a ranking of the interior vertices of \mathcal{T} corresponds to an ordering of the occurrence of the associated speciation events. An example of a ranked phylogenetic tree is illustrated in Figure 2.

For species a, b, c, d , a *relative divergence date* is a statement of the form “ $\text{div}(c, d)$ predates $\text{div}(a, b)$ ”, which is interpreted as “the divergence of c and d predates that of a and b ”. No specific dates are required to make this statement, just an ordering on the two associated speciation events.

Let \mathcal{D} be a collection of relative divergence dates. We denote the set $\bigcup \{a, b, c, d\}$ over all statements “ $\text{div}(c, d)$ predates $\text{div}(a, b)$ ” in \mathcal{D} by $\mathcal{L}(\mathcal{D})$. That is, $\mathcal{L}(\mathcal{D})$ is the set of all species that are mentioned by at least one statement of relative divergence. Also, we say that \mathcal{D} is *preserved* by a ranked phylogenetic tree (\mathcal{T}, r) with $\mathcal{L}(\mathcal{D}) \subseteq \mathcal{L}(\mathcal{T})$ if $r(c, d) < r(a, b)$ for all statements “ $\text{div}(c, d)$ predates $\text{div}(a, b)$ ” in \mathcal{D} .

In the first part of this chapter, we present an algorithm called RANKEDTREE that provides a polynomial-time solution to the following classification problem.

Problem: PHYLOGENETIC RANKING

Instance: A collection \mathcal{R} of rooted phylogenetic trees and a collection \mathcal{D} of relative divergence dates.

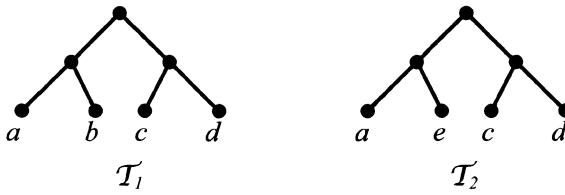


Figure 3. Two rooted phylogenetic trees \mathcal{T}_1 and \mathcal{T}_2 . Applying RANKEDTREE to these trees with the two relative divergence dates “ $\text{div}(c, d)$ predates $\text{div}(a, b)$ ” and “ $\text{div}(a, e)$ predates $\text{div}(c, d)$ ” gives the ranked phylogenetic tree shown in Figure 2.

Question: Does a ranked phylogenetic tree on $\mathcal{L}(\mathcal{R}) \cup \mathcal{L}(\mathcal{D})$ exist that displays \mathcal{R} and preserves \mathcal{D} and, if so, can we construct such a ranked phylogenetic tree in polynomial time?

To illustrate PHYLOGENETIC RANKING, consider the phylogenetic trees \mathcal{T}_1 and \mathcal{T}_2 , and the two relative divergence dates shown in Figure 3. Applying RANKEDTREE to \mathcal{T}_1 and \mathcal{T}_2 and these relative divergence dates results in the ranked phylogenetic tree (\mathcal{T}, r) shown in Figure 2. Observe that \mathcal{T} displays both \mathcal{T}_1 and \mathcal{T}_2 , and \mathcal{T} preserves both relative divergence dates. In this case, (\mathcal{T}, r) is unique; however, this does not generally happen. Note that it is possible for the interior vertices to have the same rank, in which case there is no particular ordering of the associated speciation events.

The chapter is organized as follows. In the next section, we describe briefly some necessary concepts. In Section 3, we present RANKEDTREE and show that it does indeed give a polynomial-time solution to PHYLOGENETIC RANKING. In Section 4, we extend the input of RANKEDTREE to include interval constraints on divergence dates. Lastly, in Section 5, we describe some applications and extensions of the BUILD algorithm.

2. Clusters, hierarchies, and precedence constraints

Let \mathcal{T} be a rooted phylogenetic X -tree. A *cluster* of \mathcal{T} is a subset of X whose elements are the set of descendants of a vertex of \mathcal{T} . Observe that X is a cluster of \mathcal{T} and, for all $x \in X$, $\{x\}$ is a cluster of \mathcal{T} . We denote the set of clusters of \mathcal{T} by $\mathcal{H}(\mathcal{T})$.

The set of clusters of a rooted phylogenetic X -tree is an example of a hierarchy \mathcal{H} on X ; that is, a collection of subsets of X that has the property that, for all $A, B \in \mathcal{H}$,

$$A \cap B \in \{\emptyset, A, B\}.$$

Hierarchies and rooted phylogenetic trees are related closely. In particular, given a hierarchy \mathcal{H} on X that contains X and all 1-element subsets of X , there is a unique rooted phylogenetic X -tree the set of clusters of which is \mathcal{H} . Thus, a rooted phylogenetic tree \mathcal{T} is determined by its set of clusters $\mathcal{H}(\mathcal{T})$. Indeed, \mathcal{T} can be constructed quickly and easily from $\mathcal{H}(\mathcal{T})$. Furthermore, we can associate a rank function to a hierarchy in the same way we associate a rank function to the interior vertices of a rooted phylogenetic tree. Let r be a function from $\mathcal{H} - \{\{x\} : x \in X\}$ into the set of positive integers such that, for all $A, B \in \mathcal{H} - \{\{x\} : x \in X\}$, $r(A) < r(B)$ if B is a proper subset of A . If \mathcal{H} contains X and all 1-element subsets of X , the pair (\mathcal{H}, r) is called a *ranked hierarchy* on X and the function r is a *rank function for \mathcal{H}* . Observe that, by the remarks above, we can view ranked hierarchies on X as ranked phylogenetic X -trees. This viewpoint is used freely throughout this chapter.

Lastly, a *precedence constraint* is a pairwise relationship of the form

$$(c, d) \prec (a, b),$$

where a, b, c , and d are species. Note that a, b, c, d are not necessarily different because we might wish to allow constraints such as $(a, c) \prec (a, b)$ (i.e., where the same species is involved in both divergence events). In the context of this chapter, such a constraint denotes that the divergence of c and d predates that of a and b . For a collection \mathcal{P} of precedence constraints, we denote the set $\cup\{a, b, c, d\}$ over all $(c, d) \prec (a, b) \in \mathcal{P}$ by $\mathcal{L}(\mathcal{P})$. A collection \mathcal{P} of precedence constraints is *preserved* by a ranked phylogenetic tree (\mathcal{T}, r) with $\mathcal{L}(\mathcal{P}) \subseteq \mathcal{L}(\mathcal{T})$ if $r(c, d) < r(a, b)$ for all $(c, d) \prec (a, b) \in \mathcal{P}$.

3. RANKEDTREE

In this section, we present RANKEDTREE, a polynomial-time solution to PHYLOGENETIC RANKING.

The input to RANKEDTREE is a collection of precedence constraints. These constraints are constructed from the collection \mathcal{R} of rooted phylogenetic trees and the collection \mathcal{D} of relative divergence dates. Evidently, a ranked phylogenetic tree preserves the relative divergence date “ $\text{div}(c, d)$ predates $\text{div}(a, b)$ ” if and only if it preserves $(c, d) \prec (a, b)$. We next show that a rooted phylogenetic tree \mathcal{T} displays \mathcal{R} if and only if (\mathcal{T}, r) preserves a certain collection of precedence constraints for some rank function r for \mathcal{T} .

A *rooted triple* is a rooted binary phylogenetic tree \mathcal{T} on three leaves. The rooted triple with leaves a, b , and c is denoted $ab \mid c$ if $\text{mrca}_{\mathcal{T}}(a, b)$ is a

descendant of $\text{mrca}_T(a, c)$ or, equivalently, a descendant of $\text{mrca}_T(b, c)$. The straightforward proof of the next lemma is omitted.

Lemma 3.1. *Let T be the rooted triple $ab \mid c$, let T' be a rooted phylogenetic tree with $\{a, b, c\} \subseteq L(T')$, and let r be a rank function for T' . Then T' displays T if and only if $r(a, c) < r(a, b)$.*

For a rooted phylogenetic tree T , let $\mathcal{R}(T)$ denote the set of rooted triples displayed by T . It is well known that all rooted phylogenetic trees that display $\mathcal{R}(T)$ are refinements of T (e.g., see Theorem 1 of Bryant and Steel, 1995). Lemma 3.2 is an immediate consequence of this result and Lemma 3.1.

Lemma 3.2. *Let \mathcal{R} be a collection of rooted phylogenetic trees and let*

$$\mathcal{P} = \{(a, c) \prec (a, b) : ab \mid c \in \mathcal{R}(T) \text{ and } T \in \mathcal{R}\}.$$

Let T' be a rooted phylogenetic tree with $L(T') = L(\mathcal{R})$. Then T' displays \mathcal{R} if and only if there is rank function r for T' such that (T', r) preserves \mathcal{P} .

The algorithm RANKEDTREE is shown in Figure 4. Note that, for a graph G and a subset V' of the vertex set of G , $G[V']$ denotes the subgraph of G induced by V' ; that is, the subgraph of G that has vertex set V' and edge set $\{\{a, b\} \in E(G) : a, b \in V'\}$.

Briefly, RANKEDTREE works as follows. The input to RANKEDTREE is a collection \mathcal{P} of precedence constraints. If there exists a ranked phylogenetic tree that preserves \mathcal{P} , then RANKEDTREE builds a ranked hierarchy (\mathcal{H}, r) on $L(\mathcal{P})$ recursively beginning with the hierarchy $\{L(\mathcal{P})\}$. At iteration k , a hierarchy \mathcal{H}_k on $L(\mathcal{P})$ is constructed by adding the blocks of particular partitions of minimal members of \mathcal{H}_{k-1} . Which blocks are added is determined by the components of a certain graph that is constructed at each iteration. This process ends when the constructed hierarchy contains $\{x\}$ for all $x \in L(\mathcal{P})$. If there is no such ranked phylogenetic tree, then at some iteration, i say, no new blocks are added to \mathcal{H}_{i-1} and RANKEDTREE returns “not compatible”, indicating that there is no such ranked phylogenetic tree.

Theorem 3.3 is the main result of this section.

Theorem 3.3. *Suppose that RANKEDTREE is applied to a collection \mathcal{P} of precedence constraints.*

- (i) *If RANKEDTREE returns a ranked hierarchy on $L(\mathcal{P})$, then this ranked hierarchy preserves \mathcal{P} .*

RANKEDTREE(\mathcal{P})**Input:**A collection \mathcal{P} of precedence constraints**Output:**A ranked hierarchy (\mathcal{H}, r) on $\mathcal{L}(\mathcal{P})$ that preserves \mathcal{P} or the phrase “*not compatible*” if no such hierarchy exists**Data structures:**Partitions π_1, π_2, \dots of $\mathcal{L}(\mathcal{P})$ Hierarchies $\mathcal{H}_1, \mathcal{H}_2, \dots$ on $\mathcal{L}(\mathcal{P})$ Graphs $G_1 = (\mathcal{L}(\mathcal{P}), E_1), G_2 = (\mathcal{L}(\mathcal{P}), E_2), \dots$ Rank function r

```

begin
   $k \leftarrow 1$ 
   $\mathcal{H}_k \leftarrow \{\mathcal{L}(\mathcal{P})\}$ 
   $\pi_k \leftarrow \{\mathcal{L}(\mathcal{P})\}$ 
  repeat while  $\pi_k$  contains a block with at least two elements
     $E_k \leftarrow \{(a, b) : \text{there exists } B \in \pi_k \text{ and } c, d \in B \text{ with } (c, d) \prec (a, b) \in \mathcal{P}\}$ 
     $G_k \leftarrow (\mathcal{L}(\mathcal{P}), E_k)$ 
    if  $G_k[B]$  is connected for all  $B \in \pi_k$ , then
      return “not compatible” and halt
    else
      Let  $\pi_{k+1}$  be the partition of  $X$  given by the components of  $G_k$ .
       $\mathcal{H}_{k+1} \leftarrow \mathcal{H}_k \cup \pi_{k+1}$ 
      for all blocks  $B$  in  $\pi_k$  that are not blocks of  $\pi_{k+1}$  do
         $r(B) \leftarrow k$ 
      end (for)
       $k \leftarrow k + 1$ 
    end (if-else)
  end (repeat)
   $\mathcal{H} \leftarrow \mathcal{H}_k$ 
  return  $\mathcal{H}$  and  $r$ 
end.

```

Figure 4. RANKEDTREE.

- (ii) If RANKEDTREE returns “*not compatible*”, then there is no ranked hierarchy that preserves \mathcal{P} .

Proof. To prove (i), suppose that RANKEDTREE returns a ranked hierarchy on $\mathcal{L}(\mathcal{P})$. Furthermore, suppose that $(c, d) \prec (a, b) \in \mathcal{P}$ and $r(c, d) = l$. Then c and d are elements of the same block of π_k for all $k \leq l$, and therefore $\{a, b\} \in E_k$ for all $k \leq l$. It follows that a and b are elements of the same block of π_k for all $k \leq l + 1$. Thus $r(c, d) < l + 1 \leq r(a, b)$. It follows that the returned rank hierarchy preserves \mathcal{P} .

Now consider (ii). Suppose that there exists a ranked hierarchy (\mathcal{H}^*, r^*) that preserves \mathcal{P} , but RANKEDTREE returns “not compatible” at iteration k . Then, for all blocks B of π_k , the graph $G_k[B]$ is connected.

Let B be a block of π_k that minimizes

$$\max \{r^*(A) : A \in \mathcal{H}^*, B \subseteq A\},$$

and let A be the member of \mathcal{H}^* that corresponds to this minimization. By the minimality of A , there exist disjoint members $A_1, A_2 \in \mathcal{H}^*$ such that $A_1, A_2 \subset A$, and $A_1 \cap B$ and $A_2 \cap B$ are both non-empty. Choose A_1 and A_2 to be maximal with these properties.

Because $G_k[B]$ is connected, there is an edge in this graph joining a vertex $a \in A_1 \cap B$ and a vertex $b \in A_2 \cap B$. This implies that there is a precedence constraint $(c, d) \prec (a, b) \in \mathcal{P}$ such that c and d are elements of the same block B' of π_k . Because (\mathcal{H}^*, r^*) preserves \mathcal{P} , we have $r^*(c, d) < r^*(a, b)$. By the choice of a and b , we have $r^*(a, b) = r^*(A) = r^*(B)$. But then

$$r^*(B') \leq r^*(c, d) < r^*(a, b) = r^*(B),$$

contradicting the choice of B . This completes the proof of (ii).

Proposition 3.4. *For a collection \mathcal{P} of precedence constraints, RANKEDTREE can be implemented to run in $O(|\mathcal{P}| + n^3)$ time, where $n = |\mathcal{L}(\mathcal{P})|$.*

Proof. The running time of RANKEDTREE is dominated by the time taken to complete the main loop. Provided the statement “not compatible” is not returned, each pass through this loop adds at least one cluster to the hierarchy being built. Because there are at most $2n - 2$ clusters that can be added, there are at most $O(n)$ iterations of this loop. Now, for each iteration k of this loop, we construct a graph, determine its components, and update the hierarchy assigning a ranking to certain members of the resulting hierarchy. For the first iteration, the graph G_1 takes $O(|\mathcal{P}|)$ time to construct. After that, for all k , we can construct G_{k+1} from G_k by deleting the appropriate edges. Over all iterations, the total time required to construct all these graphs is $O(|\mathcal{P}|)$. Lastly, for each iteration, it takes $O(n^2)$ time to determine the components, update the hierarchy, and assign a ranking as above.

Combining the remarks at the beginning of this section with Theorem 3.3 and Proposition 3.4, we get the following corollary immediately.

Corollary 3.5. *The algorithm RANKEDTREE provides a polynomial-time solution to PHYLOGENETIC RANKING.*



Figure 5. Two rooted phylogenetic trees T_1 and T_2 .

Remarks.

- (i) Those readers familiar with BUILD will observe that RANKEDTREE works in a similar way. Indeed, if the input to RANKEDTREE arises from just a collection \mathcal{R} of rooted phylogenetic trees and includes no relative divergence dates, then its output is identical to that returned by BUILD applied to \mathcal{R} . However, there is one significant difference. In BUILD, one considers a single minimal block of the current hierarchy at each iteration. Here, we need to consider all such blocks to guarantee an appropriate ranking.
- (ii) The rooted phylogenetic tree that is associated with the ranked phylogenetic tree returned by RANKEDTREE when applied to a collection \mathcal{R} of rooted phylogenetic trees and a collection of relative divergence dates is not necessarily a refinement of the rooted phylogenetic tree returned by BUILD when applied to \mathcal{R} . For example, consider the two rooted phylogenetic trees T_1 and T_2 shown in Figure 5 and the relative divergence date “ $\text{div}(a, c)$ predates $\text{div}(a, b)$ ”. Applying BUILD to T_1 and T_2 returns the rooted phylogenetic tree shown in Figure 6(a). However, applying RANKEDTREE to T_1 and T_2 as well as the relative divergence date, returns the ranked phylogenetic tree shown in Figure 6(b).

4. Divergence time intervals

In this section, we extend the input of RANKEDTREE to include time bounds on speciation events.

Let \mathcal{T} be a rooted phylogenetic X -tree. A *divergence time function* for \mathcal{T} is a function f from the set V° of interior vertices of \mathcal{T} into the set $\mathbf{R}^{>0}$ of positive reals, so that, if $v_1, v_2 \in V^\circ$ and v_1 is a proper ancestor of v_2 , then $f(v_1) > f(v_2)$. The pair (\mathcal{T}, f) is a *dated phylogenetic tree*. For a subset A of X of size at least two, we denote $f(\text{mrca}_{\mathcal{T}}(A))$ by $f(A)$. In the context of this

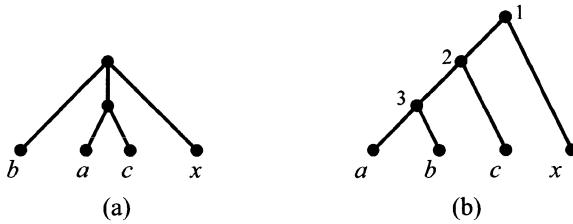


Figure 6. Rooted phylogenetic trees produced by (a) BUILD when applied to the trees \mathcal{T}_1 and \mathcal{T}_2 in Figure 5 and (b) RANKEDTREE when applied to \mathcal{T}_1 and \mathcal{T}_2 in Figure 5 as well as the relative divergence date “ $\text{div}(a, c)$ predates $\text{div}(a, b)$ ”.

chapter, the values assigned to the interior vertices of \mathcal{T} in this way represent the number of years ago that the corresponding speciation events occurred.

For species a and b , a *divergence time bound* on a and b is a lower or upper bound, denoted by $l(a, b)$ and $u(a, b)$, respectively, on the number of years ago that a and b diverged. If a and b have both a lower and upper bound, then $l(a, b) < u(a, b)$. If no information is provided concerning a lower or upper bound on the divergence time of a and b , we can always take $l(a, b) = 0$ and $u(a, b) = \infty$, respectively.

Let I be a collection of divergence time bounds. We denote the set $\cup\{a, b\}$ over all $l(a, b), u(a, b) \in I$ by $\mathcal{L}(I)$. Furthermore, I is *preserved* by a dated phylogenetic tree (\mathcal{T}, f) if $l(a, b) < f(a, b)$ and $f(a, b) < u(a, b)$ for all $l(a, b), u(a, b) \in I$.

Farach et al. (1995) showed that, given a collection I of divergence time bounds, there is a polynomial-time solution for determining and, if possible, constructing a dated phylogenetic tree with leaf set $\mathcal{L}(I)$ that preserves I . They refer to this problem as the “sandwich-to-ultrametric” problem, and the running time of the algorithm is $O(|I| + n \log(n))$ where $n = |\mathcal{L}(I)|$. In this section, we show that there is a polynomial-time solution to the following extension of this problem.

Problem: PHYLOGENETIC DIVERGENCE TIMES

Instance: A collection \mathcal{R} of rooted phylogenetic trees, a collection \mathcal{D} of relative divergence dates, and a collection I of divergence time bounds.

Question: Does a dated phylogenetic tree with leaf set $\mathcal{L}(\mathcal{R}) \cup \mathcal{L}(\mathcal{D}) \cup \mathcal{L}(I)$ exist that displays \mathcal{R} and preserves \mathcal{D} and I and, if so, can we construct such a tree in polynomial time?

Like PHYLOGENETIC RANKING, one can obtain a polynomial-time solution to PHYLOGENETIC DIVERGENCE TIMES via RANKEDTREE. Indeed,

by transforming \mathcal{I} to the collection of precedence constraints in the statement of Theorem 4.1, such a solution immediately follows from Theorems 3.3 and 4.1.

Theorem 4.1. *Let \mathcal{I} be a collection of divergence time bounds and let*

$$\mathcal{P}(l, u) = \{(c, d) \prec (a, b) : l(c, d) \geq u(a, b), \text{ where } l(c, d), u(a, b) \in \mathcal{I}\}.$$

Let \mathcal{T} be a rooted phylogenetic tree on $\mathcal{L}(\mathcal{I})$. Then there is a divergence time function f for \mathcal{T} with (\mathcal{T}, f) preserving \mathcal{I} if and only if there is a rank function r for \mathcal{T} with (\mathcal{T}, r) preserving \mathcal{P} .

Proof. First suppose that there is divergence time function f for \mathcal{T} such that (\mathcal{T}, f) preserves \mathcal{I} . Let v_1, v_2, \dots, v_n be an ordering of the interior vertices of \mathcal{T} such that

$$f(v_1) \geq f(v_2) \geq \dots \geq f(v_n).$$

Let r be the function from the set V^o of interior vertices of \mathcal{T} into the set of positive integers defined, for all i , by $r(v_i) = i$. We next show that (\mathcal{T}, r) preserves $\mathcal{P}(l, u)$.

Let $(c, d) \prec (a, b) \in \mathcal{P}(l, u)$, and let $v_i = \text{mrca}_{\mathcal{T}}(c, d)$ and $v_j = \text{mrca}_{\mathcal{T}}(a, b)$. Because $(c, d) \prec (a, b) \in \mathcal{P}(l, u)$, we have $u(a, b) \leq l(c, d)$. Therefore,

$$f(a, b) < u(a, b) \leq l(c, d) < f(c, d)$$

and so $f(v_i) > f(v_j)$. This implies that $i < j$ and so $r(c, d) = r(v_i) < r(v_j) = r(a, b)$. It follows that (\mathcal{T}, r) preserves $\mathcal{P}(l, u)$.

For the converse, suppose that there is a rank function r for \mathcal{T} such that (\mathcal{T}, r) preserves $\mathcal{P}(l, u)$. For each $v \in V^o$, let

$$(1) \quad f_l(v) = \max \{\{0\} \cup \{l(a, b) : r(a, b) \geq r(v) \text{ and } l(a, b) \in \mathcal{I}\}\}$$

and

$$(2) \quad f_u(v) = \min \{\{\infty\} \cup \{u(c, d) : r(c, d) \leq r(v) \text{ and } u(c, d) \in \mathcal{I}\}\}.$$

We show first that $f_l(v) < f_u(v)$ for all v . Suppose there exists an interior vertex v such that $f_l(v) \geq f_u(v)$. Then there are elements $a, b, c, d \in X$ such that $l(a, b) \geq u(c, d)$, $r(a, b) \geq r(v)$, and $r(c, d) \leq r(v)$. This implies that $(a, b) \prec (c, d) \in \mathcal{P}(l, u)$ and so $r(a, b) < r(c, d)$; a contradiction. Thus,

$f(v) < f_u(v)$ for all $v \in V^o$. Furthermore, by construction, if $r(v) < r(v')$ for vertices $v, v' \in V^o$, then $f(v) \geq f(v')$ and $f_u(v) \geq f_u(v')$.

Now let v_1, v_2, \dots, v_n be an ordering of the interior vertices of \mathcal{T} such that

$$r(v_1) \leq r(v_2) \leq \dots \leq r(v_n).$$

Let f be the function from V^o into the set of positive reals that is defined recursively as follows:

- (i) set $f(v_1)$ so that $f(v_1) < f(v_1) < f_u(v_1)$, and
- (ii) for all $i \in \{2, \dots, n\}$, set $f(v_i)$ so that $f(v_i) < f(v_i) < \min\{f(v_{i-1}), f_u(v_i)\}$.

Observe that, if $v = \text{mrca}_{\mathcal{T}}(a, b)$ for some $a, b \in X$, then

$$l(a, b) \leq f(v) < f(v) < f_u(v) \leq u(a, b).$$

Moreover, if v' is a proper descendent of v , then, by the minimality condition in (ii), $f(v') < f(v)$. We conclude that (\mathcal{T}, f) preserves I .

Example 4.2. To illustrate PHYLOGENETIC DIVERGENCE TIMES, suppose that we have as our instance $\mathcal{R} = \{\mathcal{T}_1, \mathcal{T}_2\}$, where \mathcal{T}_1 and \mathcal{T}_2 are as shown in Figure 7; \mathcal{D} consisting of the statement “ $\text{div}(a, e)$ predates $\text{div}(c, f)$ ”; and \mathcal{I} consisting of the divergence time bounds (in millions of years) $l(a, d) = 1$ and $u(a, d) = 3.5$, $l(a, b) = 4$ and $u(a, b) = 6$, and $l(c, f) = 3$ and $u(c, f) = 5$. Treating this instance as input, RANKEDTREE returns the ranked phylogenetic tree (\mathcal{T}, r) shown in Figure 8.

Figure 9(a) shows \mathcal{T} together with the values $f(v)$ and $f_u(v)$ for each interior vertex v given by (1) and (2), respectively. Furthermore, Figure 9(b) shows \mathcal{T} together with a divergence time function f for \mathcal{T} given by (i) and (ii) in the proof of Theorem 4.1.

Suppose that RANKEDTREE applied to collections \mathcal{R} , \mathcal{D} , and \mathcal{I} of rooted phylogenetic trees, relative divergence dates, and divergence time bounds returns a ranked phylogenetic tree. Let $a, b \in \mathcal{L}(\mathcal{R}) \cup \mathcal{L}(\mathcal{D}) \cup \mathcal{L}(\mathcal{I})$. We would like to find the *most recent* (respectively, *most ancient*) *admissible dates* of the divergence of a and b . This is the smallest (respectively, largest) date measured from the present into the past at which a and b could have diverged so that RANKEDTREE applied to these collections together with the lower bound (respectively, upper bound) corresponding to this date returns a ranked phylogenetic tree. Note that these values are not given immediately by the output of RANKEDTREE applied to \mathcal{R} , \mathcal{D} and \mathcal{I} . There are two reasons

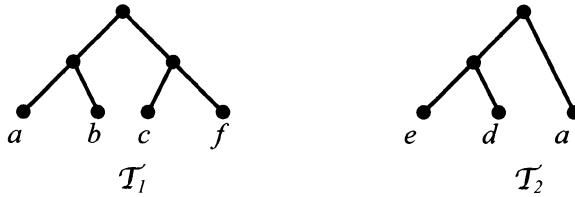


Figure 7. Two rooted phylogenetic trees T_1 and T_2 .

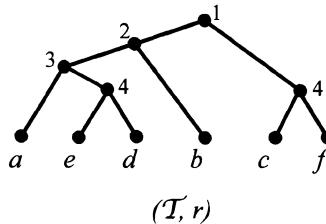


Figure 8. The ranked phylogenetic tree produced by RANKEDTREE when applied to T_1 and T_2 in Figure 7 together with the relative divergence dates and divergence time bounds given in Example 4.2.

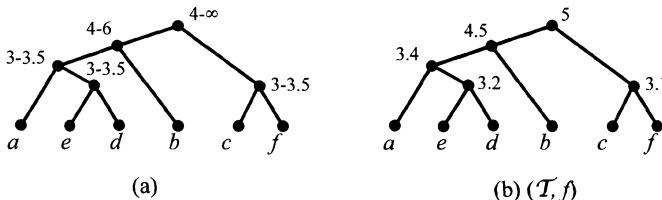


Figure 9. Assigning a divergence time function to the ranked phylogenetic tree shown in Figure 8. (a) The assignment of intervals as given by (1) and (2) in the proof of Theorem 4.1. (b) One allowable choice of divergence dates.

for this. First, for any interior vertex v , the interval $[f_l(v), f_u(v)]$ described in the proof of Theorem 4.1 does not necessarily contain all possible admissible dates for v in the tree returned by the algorithm. As a simple illustration of this, consider the pair d and e in Example 4.2. The most recent admissible divergence date for d and e is just before the present time (technically, $0 + \epsilon$, where $\epsilon > 0$) and not the value 3 as illustrated in Figure 9(a). Second, there might exist many rooted phylogenetic trees that display \mathcal{R} and preserve \mathcal{D} and \mathcal{I} for which there is a divergence time function that allows more ancient or more recent divergence times for a particular pair of species than that given by the tree returned by RANKEDTREE. To avoid having to search a possibly exponential set of trees, it is therefore comforting to have the

following result, which shows that the problem can be solved in polynomial time.

Corollary 4.3. *Suppose that RANKEDTREE applied to collections \mathcal{R} , \mathcal{D} , and \mathcal{I} returns a ranked phylogenetic tree. Let $a, b \in \mathcal{L}(\mathcal{R}) \cup \mathcal{L}(\mathcal{D}) \cup \mathcal{L}(\mathcal{I})$. Then there is a polynomial-time algorithm for determining the most ancient and most recent admissible dates for the divergence of a and b over all possible rooted phylogenetic trees and divergence time functions that display \mathcal{R} and preserve \mathcal{D} and \mathcal{I} .*

Proof. We will describe a simple polynomial algorithm that uses RANKEDTREE as a subroutine. It is possible that a more direct algorithm can be developed, but we do not explore this here.

We ensure first that u and l are defined for all pairs of species by extending \mathcal{I} as follows: if $u(x, y) \notin \mathcal{I}$, then set $u(x, y) = \infty$ and, if $l(x, y) \notin \mathcal{I}$, then set $l(x, y) = 0$.

We describe a method for determining the most ancient admissible date of the divergence of a and b ; an analogous result for most recent admissible date is similar. Let

$$T_{ab} = \{t : t \leq u(a, b) \text{ and } t = u(x, y) \text{ for some } u(x, y) \in \mathcal{I} \text{ with } u(x, y) > l(a, b)\}.$$

Then, for any $\varepsilon > 0$, the most ancient admissible date for the divergence of a and b is the maximum value of $t - \varepsilon$ over all $t \in T_{ab}$ with the property that RANKEDTREE applied to \mathcal{R} , \mathcal{D} , and \mathcal{I} together with $l(a, b) = t - \varepsilon$ returns a ranked phylogenetic tree. Clearly, the running time of this method is polynomial in $|\mathcal{L}(\mathcal{P})|$.

5. Methodological applications of BUILD

The aim of this section is to illustrate how some problems in phylogenetics that, in general, appear computationally intractable (NP-hard) can be solved efficiently in certain cases by applying the simple supertree method BUILD. We have chosen four problems to illustrate the scope of this approach. In each of these problems, the computational part of the solution is done using BUILD. The first two consider the supertree problem for unrooted phylogenetic trees and the second two consider the compatibility of two-state characters.

5.1 Combining unrooted trees

Let \mathcal{T} be a phylogenetic X -tree and let $X' \subseteq X$. Analogous to the rooted case, the *restriction of \mathcal{T} to X'* is the phylogenetic X' -tree obtained by suppressing all degree-two vertices of the minimal subtree of \mathcal{T} containing X' . We denote this restriction by $\mathcal{T}|X'$. For two phylogenetic trees \mathcal{T}_1 and \mathcal{T}_2 with $\mathcal{L}(\mathcal{T}_1) \subseteq \mathcal{L}(\mathcal{T}_2)$, we say that \mathcal{T}_2 *displays* \mathcal{T}_1 if $\mathcal{T}_2| \mathcal{L}(\mathcal{T}_1)$ is a refinement of \mathcal{T}_1 . A collection \mathcal{U} of phylogenetic trees is *compatible* if there is a phylogenetic tree \mathcal{T} that displays every member of \mathcal{U} , in which case we say that \mathcal{T} *displays* \mathcal{U} .

For a collection of phylogenetic trees, determining the compatibility of this collection is NP-complete in general (Bodlaender *et al.*, 1992; Steel, 1992). However, the first two problems show that particular cases can be solved efficiently.

5.1.1 Specified split

For our first problem, suppose we have a collection $\mathcal{U} = \{\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_k\}$ of unrooted phylogenetic trees together with a split $A|B$ of $\bigcup_{i=1}^k \mathcal{L}(\mathcal{T}_i)$. We wish to determine if there is a phylogenetic tree that displays \mathcal{U} and induces the split $A|B$. For the biologist, $A|B$ should be thought of as some “reliable” split of the taxa that is expected to be present in all acceptable supertrees of \mathcal{U} . The introduction of this split is important computationally because it can allow for the fast solution of the problem of determining the compatibility of \mathcal{U} . This is described in Theorem 5.1, the proof of which provides a polynomial-time algorithm.

Theorem 5.1. *Let $\mathcal{U} = \{\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_k\}$ be a collection of phylogenetic trees and let $A|B$ be a split of $X = \bigcup_{i=1}^k \mathcal{L}(\mathcal{T}_i)$. Suppose that, for all i , $\mathcal{L}(\mathcal{T}_i) \cap A$ and $\mathcal{L}(\mathcal{T}_i) \cap B$ are both non-empty. Then there is a polynomial-time algorithm to determine if a phylogenetic X -tree that displays \mathcal{U} and induces the split $A|B$ exists and, if so, to construct such a phylogenetic tree.*

Proof. If, for some i , the split $(A \cap \mathcal{L}(\mathcal{T}_i))|(B \cap \mathcal{L}(\mathcal{T}_i))$ is not a split of \mathcal{T}_i , then there is no phylogenetic tree satisfying the statement of the theorem. Thus, we can assume, for all i , that $\sigma_i = (A \cap \mathcal{L}(\mathcal{T}_i))|(B \cap \mathcal{L}(\mathcal{T}_i))$ is a split of \mathcal{T}_i . Now, for each i , let \mathcal{T}_i^A and \mathcal{T}_i^B denote the two rooted phylogenetic trees, with $\mathcal{L}(\mathcal{T}_i^A) = \mathcal{L}(\mathcal{T}_i) \cap A$ and $\mathcal{L}(\mathcal{T}_i^B) = \mathcal{L}(\mathcal{T}_i) \cap B$, obtained by deleting the unique edge of \mathcal{T} corresponding to σ_i and distinguishing the end vertices of this edge as root vertices. Let $\mathcal{R}^A = \{\mathcal{T}_1^A, \mathcal{T}_2^A, \dots, \mathcal{T}_k^A\}$ and $\mathcal{R}^B = \{\mathcal{T}_1^B, \mathcal{T}_2^B, \dots, \mathcal{T}_k^B\}$.

If the application of BUILD to either \mathcal{R}^A or \mathcal{R}^B detects that either of these is incompatible, then no phylogenetic X -tree exists that displays \mathcal{U} and that also induces the split $A|B$. Otherwise, we will let \mathcal{T}^A and \mathcal{T}^B denote the rooted phylogenetic trees returned by BUILD when applied to \mathcal{R}^A and \mathcal{R}^B , respectively. Thus, assume that two such trees are returned. Let \mathcal{T} be the phylogenetic X -tree obtained by adjoining the roots of \mathcal{T}^A and \mathcal{T}^B with a new edge. It now follows that \mathcal{T} displays \mathcal{U} , and $A|B$ is an induced split of \mathcal{T} .

5.1.2 Quartet compatibility

Binary phylogenetic trees with just four leaves — *quartet trees* — play a special role in phylogenetics. For example, the problem of determining the compatibility of a set of unrooted phylogenetic trees can be reduced to determining the compatibility of a set of associated quartet trees (see Steel, 1992). Furthermore, many techniques for reconstructing phylogenies (such as the “quartet puzzling” approach of Strimmer and von Haeseler, 1996) are based on quartet trees. Supertree techniques based on quartet methods have also been proposed (e.g., Piaggio-Talice *et al.*, 2004).

We will write $xy|uv$ to denote the quartet tree in which the interior edge separates the pair of leaves x, y from u, v . For a set \mathcal{Q} of quartet trees, we let

$$\mathcal{L}(\mathcal{Q}) = \bigcup_{\mathcal{T} \in \mathcal{Q}} \mathcal{L}(\mathcal{T}).$$

The proof of Theorem 5.2 describes an algorithm to determine the compatibility of an arbitrary set of quartet trees. Although not in polynomial time, this method might be practical when \mathcal{Q} is reasonably small. McMorris *et al.* (1994) provide a more general algorithm for this problem, but it has complexity $O(n^{k+1})$ where $k = |\mathcal{Q}|$ and $n = |\mathcal{L}(\mathcal{Q})|$, which, in general, will be larger than the complexity of the following approach.

Theorem 5.2. *Let \mathcal{Q} be a set of k quartet trees. Then there is an $O(k^2 2^k)$ -time algorithm for determining the compatibility of \mathcal{Q} .*

Proof. First note that, for a collection \mathcal{R} of k rooted triples, BUILD can be implemented to run on \mathcal{R} in $O(k^2)$ time (see Aho *et al.*, 1981).

To describe an algorithm that satisfies the statement of the theorem, let x be an element not in $\mathcal{L}(\mathcal{Q})$ and, for each $q = ab|cd$ in \mathcal{Q} , consider the collections of rooted triples $S_1(q) = \{cd|a, cd|b\}$ and $S_2(q) = \{ab|c, ab|d\}$. For each $q \in \mathcal{Q}$, every rooted phylogenetic tree \mathcal{T} that displays either $S_1(q)$ or $S_2(q)$ has the property that the phylogenetic tree obtained from \mathcal{T} by not distinguishing the root and suppressing this vertex if it has degree two,

displays q . Now, for each of the 2^k functions $\pi : \mathcal{Q} \rightarrow \{1, 2\}$, BUILD can determine the compatibility of $\cup_{q \in \mathcal{Q}} S_{\pi(q)}(q)$ in $O(nk)$ time. Moreover, it is checked easily that \mathcal{Q} is compatible if and only if $\cup_{q \in \mathcal{Q}} S_{\pi(q)}(q)$ is compatible for some choice of π . The theorem now follows.

We note in passing that the running time of the algorithm in Theorem 5.2 could be improved slightly by invoking the approach of Henzinger *et al.* (1999).

5.2 Two-state character compatibility

In the next two problems, we consider characters that assign one of two states to some or all of the species. More precisely, a *two-state character* on X is a function $\chi : X' \rightarrow \{0, 1\}$, where X' is some subset of X . Allowing X' to be a strict subset of X allows for uncertainty or ambiguity of the character state of certain species in X . The two-state character χ is *convex* on a phylogenetic X -tree if there exists a split $A \mid B$ of \mathcal{T} for which $\chi^{-1}(0) \subseteq A$ and $\chi^{-1}(1) \subseteq B$. Furthermore, a collection C of two-state characters on X is *compatible* if there exists a phylogenetic X -tree on which all the characters in C are convex, in which case we say that \mathcal{T} *displays* C . The biological relevance of these concepts, in particular their close connection with the concept of homoplasy, is described by Semple and Steel (2002). In general, determining the compatibility of C is an NP-complete problem.

In both problems presented here, we consider the following collection of rooted phylogenetic trees. For a two-state character $\chi : X' \rightarrow \{0, 1\}$, let \mathcal{T}_χ denote the rooted phylogenetic X' -tree that has exactly two interior vertices with the non-root vertex adjacent to the leaves in $\chi^{-1}(1)$ and the root vertex adjacent to the leaves in $\chi^{-1}(0)$. For a collection C of two-state characters on X , let

$$(3) \quad \mathcal{R}(C) = \{\mathcal{T}_\chi : \chi \in C\}.$$

5.2.1 Directed case

Let χ be a two-state character on X and let \mathcal{T} be a rooted phylogenetic X -tree. We say that χ is *convex on \mathcal{T} relative to $0 \rightarrow 1$* if there is a function $f : V(\mathcal{T}) \rightarrow \{0, 1\}$ that extends χ so that

- (i) there is no arc (u, v) of \mathcal{T} with $f(u) = 1$ and $f(v) = 0$, and
- (ii) there is at most one arc (u, v) of \mathcal{T} with $f(u) = 0$ and $f(v) = 1$.

Here, we view the edges of a rooted phylogenetic tree as arcs directed away from the root. A collection C of two-state characters on X is *compatible relative to $0 \rightarrow 1$* if there is a rooted phylogenetic X -tree \mathcal{T} on which all the characters in C are convex relative to $0 \rightarrow 1$, in which case we say that \mathcal{T} displays C relative to $0 \rightarrow 1$.

The setup of the previous paragraph is useful for modeling situations in which 0 represents some specific “ancestral” state and 1 represents a specific “derived” state, and where transitions are rare and always proceed from the ancestral to the derived state. If there is uncertainty as to whether the state for species x is ancestral or derived, no state is assigned to that species. This formulation of compatibility has been applied, for example, to certain molecular genetic data known as SINEs (“short interspersed nuclear elements”), where the states 0 and 1 denote the absence and presence, respectively, of a particular sequence inserted into a particular region of a genome (see Pe’er *et al.*, 2000). The present setting is also relevant to a modification of the MRP technique for supertree construction (Baum, 1992; Ragan, 1992) described by Bininda-Emonds and Bryant (1998), where reversals (i.e., $1 \rightarrow 0$ transitions) are prohibited during the parsimony optimization step.

The question of determining the compatibility of a collection of two-state, directed characters has been shown to have a polynomial time solution by Pe’er *et al.* (2000), and, as a special case of a more general result, by Benham *et al.* (1995). The following theorem provides a further polynomial-time approach to the problem, showing that it can be regarded as a special case of the supertree problem for rooted phylogenetic trees.

Theorem 5.3.

- (i) Let $\chi : X' \rightarrow \{0, 1\}$ be a two-state character on X and let \mathcal{T} be a rooted phylogenetic X -tree. Then χ is convex on \mathcal{T} relative to $0 \rightarrow 1$ if and only if \mathcal{T} displays \mathcal{T}_χ .
- (ii) Let C be a collection of two-state characters on X . Then a rooted phylogenetic X -tree \mathcal{T} displays C relative to $0 \rightarrow 1$ if and only if \mathcal{T} displays $\mathcal{R}(C)$

Proof. Now χ is convex on \mathcal{T} relative to $0 \rightarrow 1$ if and only if there exists a cluster A of \mathcal{T} such that

$$(4) \quad \chi^{-1}(1) \subseteq A \text{ and } \chi^{-1}(0) \subseteq X - A.$$

Furthermore, (4) holds if and only if $\chi^{-1}(1)$ is a cluster of $\mathcal{T}|X'$. Because $\chi^{-1}(1)$ is the only -non-trivial cluster of \mathcal{T}_χ , this last condition holds if and

only if \mathcal{T} displays \mathcal{T}_χ . This completes the proof of (i). Part (ii) is an immediate consequence of (i).

5.2.2 Undirected case

In this problem, we again consider two-state characters taking values in the set $\{0, 1\}$ except that we no longer regard 0 as “ancestral”. The notion of convexity for a character on a phylogenetic tree \mathcal{T} as defined before the previous problem is consistent with the concept of characters evolving without homoplasy — it allows at most one occurrence of either the transition $0 \rightarrow 1$ or the transition $1 \rightarrow 0$ on \mathcal{T} .

Theorem 5.4. *Let $C = \{\chi_1, \chi_2, \dots, \chi_k\}$ be a collection of two-state characters on X . Suppose that, for all $i, j \in \{1, 2, \dots, k\}$,*

$$\chi_i^{-1}(0) \cap \chi_j^{-1}(0) \neq \emptyset.$$

Then there is a polynomial-time algorithm to determine if C is compatible and, if so, to construct a phylogenetic X -tree that displays C .

Proof. We establish Theorem 5.4 by showing that C is compatible if and only if the associated collection $\mathcal{R}(C)$ of rooted phylogenetic trees (described by (3)) is compatible, and that, when this occurs, the rooted phylogenetic tree returned by BUILD when applied to $\mathcal{R}(C)$ immediately gives a phylogenetic tree that displays C .

First, suppose that $\mathcal{R}(C)$ is compatible and that \mathcal{T} is a rooted phylogenetic X -tree that displays $\mathcal{R}(C)$ (e.g., the rooted phylogenetic tree produced by BUILD when applied to $\mathcal{R}(C)$). Let \mathcal{T}'^ρ be the phylogenetic X -tree obtained from \mathcal{T} by not distinguishing the root ρ . Note that if ρ has degree two, then this vertex is also suppressed. By Theorem 5.3, for each $\chi \in C$, χ is convex on \mathcal{T} relative to $0 \rightarrow 1$. But this implies that χ is convex on \mathcal{T}'^ρ . It follows that \mathcal{T}'^ρ displays C .

Now suppose that C is compatible and that \mathcal{T} is a phylogenetic X -tree that displays C . For all $i \in \{1, 2, \dots, k\}$, let V_i denote the vertex set of the minimal subtree of \mathcal{T} that contains the leaves of $\chi_i^{-1}(0)$. Because $\chi_i^{-1}(0) \cap \chi_j^{-1}(0) \neq \emptyset$, for all $i, j \in \{1, 2, \dots, k\}$, it follows that

$$V_i \cap V_j \neq \emptyset$$

for all i, j . Thus, by the Helly intersection property of subtrees of a tree (see Golumbic, 1980), there exists a vertex v_ρ of \mathcal{T} such that

$$v_\rho \in \bigcap_{i=1}^k V_i.$$

If v_ρ is an interior vertex of \mathcal{T} , then it is checked easily that the rooted phylogenetic X -tree obtained by rooting \mathcal{T} on v_ρ displays $\mathcal{R}(C)$. If v_ρ is not an interior vertex, then the rooted phylogenetic tree obtained by rooting \mathcal{T} on the vertex adjacent to v_ρ displays $\mathcal{R}(C)$. In both cases, the resulting rooted phylogenetic X -tree displays $\mathcal{R}(C)$. This completes the proof of the theorem.

Note that a particular case where the intersection condition in Theorem 5.4 applies is when

$$|\chi_i^{-1}(0)| > \frac{1}{2}|X|$$

for all $i \in \{1, 2, \dots, k\}$. In particular, we have the following corollary.

Corollary 5.5. *Let C be a collection of two-state characters on X . If each character in C assigns a strict majority of elements of X to some particular state, then there is a polynomial-time algorithm for determining the compatibility of C .*

6. Conclusion

Supertree methods continue to be extended and applied in various ways to study interesting problems in phylogenetics. In this chapter, we have considered how the BUILD algorithm can be extended to account for relative and absolute divergence date information. In the last section, we also described applications of BUILD to some other problems arising in phylogeny reconstruction.

One of the limitations of these approaches is that they are essentially “all-or-nothing”; that is, they return a tree (and dates) only if the input trees (and input dates) are compatible. But, incompatibility tends to be the rule rather than the exception for real biological data. However, the point in developing “exact” methods such as BUILD and the extensions described in this paper is that they form the basis for methods that apply in the general (incompatible) setting. Indeed, the BUILD approach was extended recently to the supertree method MINCUTSUPERTREE (Semple and Steel, 2000; also Page, 2002) that returns a supertree for every input of rooted phylogenetic trees. We believe that similar techniques can be applied to extend algorithms such as RANKEDTREE, and we hope to explore this in further work.

Acknowledgements

We thank the New Zealand Marsden Fund (UOC-MIS-005) for supporting this research, and Mike Charleston, Olaf Bininda-Emonds and an anonymous referee for helpful comments on an earlier version of this chapter.

References

- AHO, A. V., SAGIV, Y., SZYMANSKI, T. G., AND ULLMAN, J. D. 1981. Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. *SIAM Journal on Computing* 10:405–421.
- BARTHÉLEMY, J.-P. AND GUÉNOCHE, A. 1991. *Trees and Proximity Representations*. John Wiley and Sons, United Kingdom.
- BAUM, B. R. 1992. Combining trees as a way of combining datasets for phylogenetic inference, and the desirability of combining gene trees. *Taxon* 41:3–10.
- BENHAM, C., KANNAN, S., PATERSON, M., AND WARNOW, T. 1995. Hen's teeth and whale's feet: generalized characters and their compatibility. *Journal of Computational Biology* 2:515–525.
- BININDA-EMONDS, O. R. P. AND BRYANT H. N. 1998. Properties of matrix representation with parsimony analyses. *Systematic Biology* 47:497–508.
- BODLAENDER, H. L., FELLOWS, M. R., AND WARNOW, T. J. 1992. Two strikes against perfect phylogeny. In *Proceedings of the International Colloquium on Automata, Languages and Programming*, Lecture Notes in Computer Science, volume 623, pp. 273–283. Springer-Verlag, Berlin.
- BRYANT, D. AND STEEL, M. 1995. Extension operations on sets of leaf-labelled trees. *Advances in Applied Mathematics* 16:425–453.
- CONSTANTINESCU, M. AND SANKOFF, D. 1995. An efficient algorithm for supertrees. *Journal of Classification* 12:101–112.
- FARACH, M., KANNAN, S., AND WARNOW, T. 1995. A robust model for finding optimal evolutionary trees. *Algorithmica* 13:155–179.
- GOLUMBIC, M. C. 1980. *Algorithmic Graph Theory and Perfect Graphs*. Academic Press, New York.
- HENZINGER, M. R., KING, V., AND WARNOW, T. 1999. Constructing a tree from homeomorphic subtrees, with applications to computational evolutionary biology. *Algorithmica*, 24:1–13.
- MCMORRIS, F. R., WARNOW, T. J., AND WIMER, T. 1994. Triangulating vertex-colored graphs. *SIAM Journal on Discrete Mathematics* 7:296–306.
- NG, M. P. AND WORMALD, N. C. 1996. Reconstruction of rooted trees from subtrees. *Discrete Applied Mathematics*, 69:19–31.
- PAGE, R. D. M. 2002. Modified mincut supertrees. In R. Guigó and D. Gusfield (eds), *Proceedings of the Second International Workshop on Algorithms in Bioinformatics WABI 2002*, pp. 537–552, Springer-Verlag, New York.
- PE’ER, I., SHAMIR, R., AND SHARAN, R. 2000. Incomplete directed perfect phylogeny. In D. Sankoff (ed.), *Proceedings of the Eleventh Symposium on Combinatorial Pattern Matching CPM*, Lecture Notes in Computer Science 1848:143–153. Springer, New York.

- PIAGGIO-TALICE, R., BURLEIGH, J. G., AND EULENSTEIN, E. 2004. Quartet supertrees. In O. R. P. Bininda-Emonds (ed). *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 173–191. Kluwer Academic, Dordrecht, the Netherlands.
- RAGAN, M. A. 1992. Phylogenetic inference based on matrix representation of trees. *Molecular Phylogenetics and Evolution* 1:53–58.
- SEMPLE, C. 2003. Reconstructing minimal rooted trees. *Discrete Applied Mathematics* 127:489–503.
- SEMPLE, C. AND STEEL, M. 2000. A supertree method for rooted trees. *Discrete Applied Mathematics* 105:147–158.
- SEMPLE, C. AND STEEL, M. 2002. Tree reconstruction from multi-state characters. *Advances in Applied Mathematics* 28:169–184.
- SEMPLE, C. AND STEEL, M. 2003. *Phylogenetics*. Oxford University Press, Oxford.
- STEEL, M. 1992. The complexity of reconstructing trees from qualitative characters and subtrees. *Journal of Classification* 9:91–116.
- STRIMMER, K. AND VON HAESELER, A. 1996. Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. *Molecular Biology and Evolution* 13:964–969.

Chapter 7

SUPERTREE ALGORITHMS FOR NESTED TAXA

Philip Daniel and Charles Semple

Abstract: Most supertree algorithms combine collections of rooted phylogenetic trees with overlapping leaf sets into a single rooted phylogenetic tree. It is implicit in all these algorithms that the leaves of the rooted phylogenetic trees in the input collection, as a whole, represent non-nested taxa. Thus, for example, the “domestic dog” and “mammal” cannot be represented by two distinct leaves in such a collection because the former is nested inside the latter. In practice, however, one often wants to combine rooted trees in which taxa are nested. In other words, to combine rooted trees in which the leaves as well as some of the interior vertices are labeled. These interior labels represent taxa at a level higher than that of their descendants (e.g., families versus genera, or genera versus species). Moreover, it could happen that a leaf of one of the input trees represents a taxon that is represented by an interior label of another tree. In this chapter, we describe two supertree algorithms for combining rooted trees in which the leaves as well as some of the interior vertices are labeled. Called “rooted semi-labeled trees”, these trees are more general than rooted phylogenetic trees in that not only are their leaves labeled, but some of their interior vertices might be as well. Both algorithms are polynomial-time in the size of the input and are motivated by a problem posed by Page in an earlier chapter called “Taxonomy, Supertrees, and the Tree of Life”.

Keywords: BUILD; interior labels; leaf labels; nested taxa; taxonomy

1. Introduction

Throughout the chapter, we will assume that the reader is familiar with the basics of phylogenetic trees.

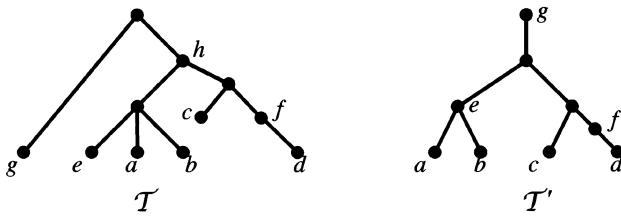


Figure 1. Two rooted semi-labeled trees.

A *rooted semi-labeled tree* (on X) is an ordered pair $(T; \phi)$ consisting of a rooted tree T with vertex set V and root vertex ρ , and a map $\phi : X \rightarrow V$ with the properties that, for all $v \in V - \{\rho\}$ of degree at most two, $v \in \phi(X)$ and, if ρ has degree zero or one, $\rho \in \phi(X)$. Viewing a rooted tree with edges directed away from the root, every vertex of out-degree 0 or 1 is labeled by an element of X . Intuitively, a rooted semi-labeled tree is simply a rooted tree in which the leaves and some of the interior vertices are labeled. We could avoid some of the technicalities of the formal definition, but then some of the generality is also lost. Rooted semi-labeled trees on X are also called *rooted X -trees*. Two rooted semi-labeled trees are shown in Figure 1. Rooted X -trees extend the notion of rooted phylogenetic X -trees; in the case of the latter, ϕ is a bijective map from X into the set of leaves of T , and the root has degree at least two.

Let $\mathcal{T} = (T; \phi)$ be a rooted X -tree and let X' be a subset of X . The *restriction* of \mathcal{T} to X' , denoted $\mathcal{T}|X'$, is the rooted X' -tree that is obtained from the minimal rooted subtree of T induced by the elements of the set $\phi(X')$ by suppressing all vertices of out-degree 0 or 1 not in $\phi(X')$. We say a rooted X -tree \mathcal{T} *displays* a rooted X' tree \mathcal{T}' if $X' \subseteq X$ and $\mathcal{T}|X'$ is a refinement of \mathcal{T}' . For example, in Figure 1, \mathcal{T} displays \mathcal{T}' . A collection \mathcal{P} of rooted semi-labeled trees is *compatible* if there exists a rooted semi-labeled tree \mathcal{T} that displays every tree in \mathcal{P} , in which case, \mathcal{T} displays \mathcal{P} .

One of the first supertree methods is due to Aho *et al.* (1981). Intended originally for relational databases, their method (called BUILD) provides a polynomial-time solution to the following problem: given a collection \mathcal{P} of rooted phylogenetic trees, does there exist a rooted phylogenetic tree that displays \mathcal{P} and, if so, can we construct such a rooted phylogenetic tree? In terms of evolutionary biology, where one views a rooted phylogenetic tree as representing the ancestral relationships of a set of present-day species, a rooted phylogenetic tree \mathcal{T} displays \mathcal{P} if, up to “polytomies”, all the ancestral relationships of every tree in \mathcal{P} is preserved in \mathcal{T} . As noted in Semple and Steel (2003), BUILD can be extended easily to determine the compatibility of a collection \mathcal{P}' of rooted semi-labeled trees in polynomial time, and, if they

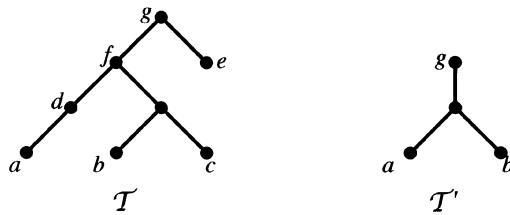


Figure 2. T perfectly displays T' .

are compatible, to construct a rooted semi-labeled tree that displays this collection. However, under this notion of compatibility, it is possible that \mathcal{P}' is compatible, but the only rooted semi-labeled trees that display \mathcal{P}' all have a particular label labeling a leaf that labeled an interior vertex of a tree in \mathcal{P}' originally. Hence, this notion of compatibility does not preserve descendancy and so, for practical purposes, it appears that it might not be of much use for collections of rooted semi-labeled trees.

In this chapter, we present two polynomial-time supertree algorithms for rooted semi-labeled trees. The first algorithm provides a solution to a problem posed by Page (2004) in another chapter of this book. We call this problem HIGHER TAXA COMPATIBILITY. A rooted X -tree T *perfectly displays* a rooted X' -tree T' if $X' \subseteq X$ and $T|_{X'}$ is isomorphic to T' . In Figure 2, T perfectly displays T' . Furthermore, a collection \mathcal{P} of rooted semi-labeled trees is *perfectly compatible* if there exists a rooted semi-labeled tree T that perfectly displays every tree in \mathcal{P} , in which case, T *perfectly displays* \mathcal{P} . Note that if T perfectly displays T' , then T displays T' . However, the converse does not hold necessarily.

Problem: HIGHER TAXA COMPATIBILITY

Instance: A collection \mathcal{P} of rooted semi-labeled trees.

Question: Does there exist a rooted semi-labeled tree that perfectly displays \mathcal{P} and, if so, can we construct such a rooted semi-labeled tree?

The motivation for HIGHER TAXA COMPATIBILITY arises when one wants to use a supertree method for combining evolutionary trees the internal vertices of which as well as their leaves are labeled. Such trees contain taxa at different taxonomic levels. Indeed, it could happen that a higher taxon labels an internal vertex in one of the trees we want to combine, but it labels a leaf in another. The algorithm that we present to solve this problem is called SEMI-LABELEDBUILD.

Our first response in trying to solve HIGHER TAXA COMPATIBILITY was to use the extension of BUILD for determining the compatibility of a collection of rooted semi-labeled trees in some way. However, despite SEMI-LABELEDBUILD having a similar description to this extension, it seems that no straightforward modification solves this problem.

The notion of perfectly displays is very restrictive in the sense that a collection \mathcal{P} of semi-labeled trees is perfectly compatible precisely if there is a rooted semi-labeled tree that preserves all the most recent common ancestor relationships described by \mathcal{P} . Thus, perfectly displays does not allow for the resolution of any polytomies in \mathcal{P} . To accommodate the possible resolution of polytomies, but still preserve all the descendants relationships in \mathcal{P} , we introduce a second notion of displays called “ancestrally displays” in the second half of this chapter. In comparison with the other two notions of displays, ancestrally displays is weaker than perfectly displays because it might not preserve all the most recent common ancestor relations, but it is stronger than the usual notion of displays because it preserves descendants. The second algorithm in this chapter determines the compatibility of \mathcal{P} under this second notion.

Unless otherwise stated, the notation and terminology in this chapter follow Semple and Steel (2003). The chapter is organized as follows. In Section 2, we describe some necessary preliminaries. In Section 3, we present SEMI-LABELEDBUILD and show that it does indeed provide a polynomial-time solution to HIGHER TAXA COMPATIBILITY. The last section defines ancestrally displays formally and presents a polynomial-time algorithm for solving the associated compatibility problem.

2. Preliminaries

Let $\mathcal{T} = (T; \phi)$ be a rooted semi-labeled tree. The tree T and the map ϕ are called the *underlying tree* and *labeling map* of \mathcal{T} . The domain of ϕ is the *label set* of \mathcal{T} and is denoted $\mathcal{L}(\mathcal{T})$. We shall often refer to the elements of $\mathcal{L}(\mathcal{T})$ as *labels*. If v is a vertex of T , we say that the elements of $\phi^{-1}(v)$ *label* v . If ρ is the root of T , then the elements of $\phi^{-1}(\rho)$ are called *root labels*. Furthermore, T is *fully labeled* if $\phi^{-1}(v)$ is non-empty for all vertices v of T . For a collection \mathcal{P} of rooted semi-labeled trees, we denote the set $\bigcup_{\mathcal{T} \in \mathcal{P}} \mathcal{L}(\mathcal{T})$ by $\mathcal{L}(\mathcal{P})$.

For a rooted tree T , a particularly useful partial order \leq_T on the vertex set V of T is obtained by setting $u \leq_T v$ if the path from the root of T to v includes u . If $u \leq_T v$, we say that v is a *descendant* of u or, alternatively, u is an *ancestor* of v . Furthermore, u and v are *comparable* under \leq_T if either $u \leq_T v$ or $v \leq_T u$; otherwise u and v are *not comparable*. Observe that the partial

order \leq_T has the property that, for every pair of elements, the greatest lower bound exists. The greatest lower bound of x and y under \leq_T is called the *most recent common ancestor* of x and y and is denoted $\text{mrca}_T(x, y)$.

The above partial order extends naturally to the label set of a rooted semi-labeled tree as follows. Let $\mathcal{T} = (T; \phi)$ be a rooted X -tree and let $a, b \in X$. Then, $a \leq_T b$ if $\phi^{-1}(a) \leq_T \phi^{-1}(b)$, in which case, b is a *descendant label* of a or, alternatively, a is an *ancestor label* of b . Furthermore, for all $a, b \in X$, we let

$$\text{mrca}_T(a, b) = \phi^{-1}(\text{mrca}_T(\phi(a), \phi(b))).$$

Note that, because \mathcal{T} is only semi-labeled, this set could be empty. However, if \mathcal{T} is fully-labeled, then this set is non-empty.

Let $\mathcal{T} = (T; \phi)$ be a rooted semi-labeled tree and let u be a vertex of T . An element a of $L(\mathcal{T})$ is a *descendant label* of u if $u \leq_T \phi^{-1}(a)$. The set of descendant labels of a non-root vertex v of T is called a *cluster* and we often refer to it as the *cluster of T corresponding to v* . The collection of clusters of \mathcal{T} is denoted by $\mathcal{H}(\mathcal{T})$. Up to isomorphism of the underlying trees, no two rooted semi-labeled trees have the same collection of clusters; thus, a rooted semi-labeled tree \mathcal{T} is determined completely by its set of clusters (see Theorem 3.5.2 in Semple and Steel, 2003). Indeed, \mathcal{T} can be constructed quickly and easily from $\mathcal{H}(\mathcal{T})$.

In Section 1, we defined the restriction of a rooted X -tree \mathcal{T} to a subset X' of X and denoted it by $\mathcal{T}|X'$. An equivalent definition of $\mathcal{T}|X'$ is the rooted X' -tree for which

$$\mathcal{H}(\mathcal{T}|X') = \{C \cap X' : C \in \mathcal{H}(\mathcal{T}) \text{ and } C \cap X' \neq \emptyset\}$$

This equivalence will prove useful in this chapter.

3. The SEMI-LABELEDBUILD algorithm

In this section, we describe the algorithm SEMI-LABELEDBUILD. We begin by defining a particular graph that will play a prominent role in the algorithm. A vertex of a graph is *isolated* if it is incident with no edges. Let \mathcal{P} be a collection of rooted fully-labeled trees. The *cluster and root-label* graph of \mathcal{P} , and denoted $G(\mathcal{P})$, has vertex set $L(\mathcal{P})$ and an edge set consisting of three types of edges that are added sequentially as follows:

- (i) First, two (not necessarily distinct) vertices are joined by a blue edge if they appear in the same cluster of a fully-labeled tree in \mathcal{P} . Note

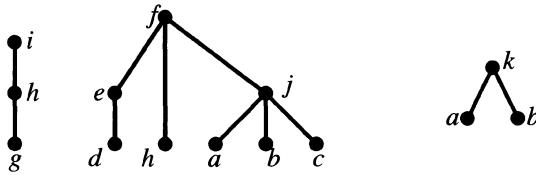


Figure 3. A collection $G(\mathcal{P})$ of rooted fully-labeled trees.

- this implies that, for any element of $\mathcal{L}(\mathcal{P})$ that is in a cluster, there is a loop joining this element to itself.
- (ii) Second, if a is a root label of a fully-labeled tree, \mathcal{T} say, in \mathcal{P} and a is not isolated, then join a to every other label in $\mathcal{L}(\mathcal{T})$ by a blue edge.
 - (iii) Third, once all of the edges associated with (i) and (ii) have been added, do the following:
 - (a) For all isolated vertices c , if there is a tree \mathcal{T} in \mathcal{P} with labels $a, b \in \mathcal{L}(\mathcal{T})$ such that $c \in \text{mrca}_{\mathcal{T}}(a, b)$, join a and b with a red edge labeled c . This labeled edge can be in parallel with a blue edge or other red edges (see remark below).
 - (b) For any two vertices joined by a red edge labeled c , if there is a path connecting them consisting of just blue edges, delete every red edge labeled c and join c to each $d \in \mathcal{L}(\mathcal{P})$ with a blue edge if both c and d are in the label set of some particular tree in \mathcal{P} .
 - (c) Repeat (b) until there are no pairs of vertices joined by red edges that are connected by a path consisting of just blue edges.
 - (d) Delete all remaining red edges.

We call a blue edge of $G(\mathcal{P})$ a *type (i), (ii), or (iii)* edge depending upon whether it is added at Step (i), (ii), or (iii), respectively, in the construction of $G(\mathcal{P})$.

Remark. In the construction of $G(\mathcal{P})$, once a blue edge has been added to join two, not necessarily distinct, vertices, no further blue edge need be added in parallel with this edge. However, red edges with distinct labels are added in parallel because each such edge plays a particular role in the construction.

As an example of the above construction, let \mathcal{P} be the collection of rooted fully-labeled trees shown in Figure 3. Then, omitting loops, the construction of $G(\mathcal{P})$ to the end of Step (ii) is shown in Figure 4. At Step (iii)(a), red edges are added between pairs $\{a, b\}, \{d, a\}, \{d, b\}, \{d, c\}, \{d, j\}, \{e, a\}, \{e, b\}, \{e, c\}, \{e, j\}, \{h, a\}, \{h, b\}, \{h, c\}, \{h, j\}, \{h, d\}$ and $\{h, e\}$. The red

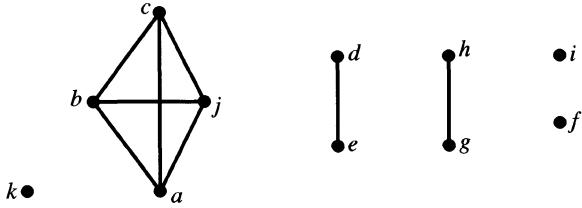


Figure 4. Construction of $G(\mathcal{P})$ at the end of steps (i) and (ii).

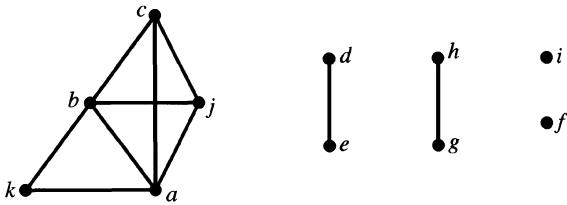


Figure 5. Completed construction of $G(\mathcal{P})$.

edge between a and b gives blue edges between k and a and between k and b at Step (iii)(b). All other red edges are deleted at Step (iii)(d) without having affected the graph. With loops omitted, the final construction of $G(\mathcal{P})$ is shown in Figure 5.

Before describing SEMI-LABELEDBUILD, we need to define one further construction. Let $\mathcal{T} = (T; \phi)$ be a rooted semi-labeled tree on X , where T has vertex set V . We say that a rooted fully-labeled tree $\mathcal{T}_1 = (T; \phi_1)$ on X_1 , where $X \subseteq X_1$, has been obtained from \mathcal{T} by *adding distinct new labels* if, for all distinct $u, v \in V$, the following properties are satisfied:

1. If $\phi^{-1}(v)$ is non-empty, then $\phi_1^{-1}(v) = \phi^{-1}(v)$.
2. If $\phi^{-1}(v)$ is empty, then $|\phi_1^{-1}(v)| = 1$.
3. If $\phi^{-1}(u)$ and $\phi^{-1}(v)$ are both empty, then $\phi_1^{-1}(u) \neq \phi_1^{-1}(v)$.

Intuitively, \mathcal{T}_1 has been obtained from \mathcal{T} by labeling non-labeled vertices of \mathcal{T} singularly with distinct new labels. For a collection \mathcal{P} of rooted semi-labeled trees, we say that \mathcal{P}_1 has been obtained from \mathcal{P} by *adding distinct new labels* if it has been obtained by adding distinct new labels to every tree in \mathcal{P} so that, for any pair of trees, no two new labels are the same.

Essentially, all the work in SEMI-LABELEDBUILD is done by a subroutine called FULLY-LABELEDBUILD. We describe each in turn.

Algorithm: SEMI-LABELEDBUILD(\mathcal{P}, \mathcal{T})**Input:** A collection \mathcal{P} of rooted semi-labeled trees.**Output:** A rooted semi-labeled tree \mathcal{T} that perfectly displays \mathcal{P} or the statement *not perfectly compatible*.

1. Construct a collection \mathcal{P}' of rooted fully-labeled trees from \mathcal{P} by adding distinct new labels.
2. Call the subroutine FULLY-LABELEDBUILD($\mathcal{P}', v', \mathcal{T}'$).
3. If FULLY-LABELEDBUILD returns *not perfectly compatible*, then return *not perfectly compatible*.
4. If FULLY-LABELEDBUILD returns a rooted fully-labeled tree \mathcal{T}' , then remove the added labels and return the resulting rooted semi-labeled tree \mathcal{T} .

Algorithm: FULLY-LABELEDBUILD($\mathcal{P}', v', \mathcal{T}'$)**Input:** A collection \mathcal{P}' of rooted fully-labeled trees.**Output:** A rooted fully-labeled tree \mathcal{T}' with root vertex v' that perfectly displays \mathcal{P}' or the statement *not perfectly compatible*.

1. Construct the cluster and root-label graph $G(\mathcal{P}')$ of \mathcal{P}' .
2. If $G(\mathcal{P}')$ has no isolated vertices, then halt and return *not perfectly compatible*.
3. Otherwise, let S_1, S_2, \dots, S_k denote the vertex sets of the connected components of $G(\mathcal{P}')$ not consisting of an isolated vertex, and let S_0 denote the set of isolated vertices of $G(\mathcal{P}')$.
4. Initialize \mathcal{T}' with a single root vertex v' and assign all labels in S_0 to v' .
5. For each $i \in \{1, 2, \dots, k\}$, call FULLY-LABELEDBUILD($\mathcal{P}', v', \mathcal{T}'$), where \mathcal{P}'_i is the collection of rooted fully-labeled trees obtained from \mathcal{P}' by restricting each tree in \mathcal{P}' to S_i . If FULLY-LABELEDBUILD($\mathcal{P}', v', \mathcal{T}'$) returns a tree, then attach \mathcal{T}'_i to v' via the edge $\{v', v'\}_i$.

Intuitively, for a set \mathcal{P}' of rooted fully-labeled trees, FULLY-LABELEDBUILD attempts to construct a rooted fully-labeled tree \mathcal{T}' that perfectly displays \mathcal{P}' by essentially constructing $\mathcal{H}(\mathcal{T}')$. This is done by beginning with $\mathcal{L}(\mathcal{P}')$ and breaking it down successively into disjoint subclusters. How clusters are broken up in this way is determined by the connected components of the associated cluster and root-label graph. Components consisting of isolated vertices are distinguished from those not consisting of isolated vertices. This process continues provided the

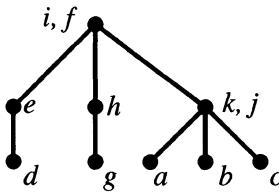


Figure 6. The rooted fully-labeled tree outputted by FULLY-LABELEDBUILD when applied to the collection of trees shown in Figure 3.

associated cluster and root-label graph has at least one isolated vertex at each iteration, in which case, FULLY-LABELEDBUILD returns a rooted fully-labeled tree. By contrast, if the associated cluster and root-label graph has no isolated vertices at some iteration, then FULLY-LABELEDBUILD returns “ \mathcal{P}' is not perfectly compatible”.

Remark.

1. It is an immediate consequence of Lemma 3.2 below that, for all i in Step 4 of FULLY-LABELEDBUILD, \mathcal{P}'_i is indeed a collection of rooted fully-labeled trees as indicated in this step.
2. Using the fact that FULLY-LABELEDBUILD considers proper restrictions of the input collection of rooted fully-labeled trees successively in Step 4 of the algorithm, it is seen easily that FULLY-LABELEDBUILD returns either “ \mathcal{P}' is not perfectly compatible” or a rooted fully-labeled tree with label set $\mathcal{L}(\mathcal{P}')$. Consequently, SEMI-LABELEDBUILD returns either “ \mathcal{P} is not perfectly compatible” or a rooted semi-labeled tree with label set $\mathcal{L}(\mathcal{P})$.

To illustrate FULLY-LABELEDBUILD, the rooted fully-labeled tree shown in Figure 6 is the result of applying this algorithm to the collection of rooted fully-labeled trees shown in Figure 3.

The main result of this chapter is the following theorem.

Theorem 3.1. *Let \mathcal{P} be a collection of rooted semi-labeled trees. Then SEMI-LABELEDBUILD applied to \mathcal{P} either:*

- (i) *returns a rooted semi-labeled tree that perfectly displays \mathcal{P} if \mathcal{P} is perfectly compatible, or*
- (ii) *returns the statement \mathcal{P} is not perfectly compatible otherwise.*

To prove Theorem 3.1, we establish first some lemmas.

Lemma 3.2. Let \mathcal{P} be a collection of rooted fully-labeled trees and consider the cluster and root-label graph $G(\mathcal{P})$ of \mathcal{P} . Let T be an element of \mathcal{P} and let S_0 denote the set of isolated vertices of $G(\mathcal{P})$. Then the following hold:

- (i) If $\mathcal{L}(P) \cap S_0$ is non-empty, then all the elements of this set are root labels of T . Furthermore, if d is a root label of T and $d \in S_0$, then all root labels of T are elements of S_0 .
- (ii) If no root label of T is an element of S_0 , then $\mathcal{L}(T)$ is a subset of the vertex set of some connected component of $G(\mathcal{P})$.
- (iii) Suppose that a root label of T is an element of S_0 . Let A and B be distinct maximal clusters of T . Then A and B are subsets of the vertex sets of distinct connected components of $G(\mathcal{P})$.

Proof. Using the construction of $G(\mathcal{P})$, the proofs of (i), (ii), and (iii) are straightforward. We omit the details.

Lemma 3.3. Let T_1 be a rooted fully-labeled tree on X_1 and let T_2 be a rooted semi-labeled tree on X_2 such that $X_1 \subseteq X_2$. Then T_2 perfectly displays T_1 if and only if, for all $a, b \in X_1$,

$$\text{mrca}_{T_1}(a, b) = \text{mrca}_{T_2}(a, b) | X_1.$$

Proof. For each $i \in \{1, 2\}$, let \mathcal{H}_i denote the collection of clusters of T_i plus X_i . Suppose that T_2 perfectly displays T_1 . Let $a, b, c \in X_1$ and suppose that $c \in \text{mrca}_{T_1}(a, b)$. Then, for any $C_1 \in \mathcal{H}_1$, $\{a, b\}$ is a subset of C_1 if and only if $c \in C_1$. Because $\mathcal{H}_1 = \mathcal{H}_2 | X_1$, it follows that, for any $C_2 \in \mathcal{H}_2$, $\{a, b\}$ is a subset of C_2 if and only if $c \in C_2$. Therefore, c is a common ancestor of a and b in T_2 , and no common ancestor of a and b in T_2 is a proper descendant of c in T_2 . Thus, $c \in \text{mrca}_{T_2}(a, b) | X_1$, and, more generally, $\text{mrca}_{T_1}(a, b) \subseteq \text{mrca}_{T_2}(a, b) | X_1$. For all x in $X_1 - \text{mrca}_{T_1}(a, b)$, there is either an element of \mathcal{H}_1 containing x but not c , or there is an element of \mathcal{H}_1 containing c but not x . Thus, because $\mathcal{H}_1 = \mathcal{H}_2 | X_1$, there is either an element of \mathcal{H}_2 containing x but not c or there is an element of \mathcal{H}_2 containing c but not x . Therefore, c and x do not label the same vertex of T_2 , and so $x \notin \text{mrca}_{T_2}(a, b) | X_1$. This shows that $\text{mrca}_{T_1}(a, b) = \text{mrca}_{T_2}(a, b) | X_1$.

Now suppose that, for all $a, b \in X_1$, we have $\text{mrca}_{T_1}(a, b) = \text{mrca}_{T_2}(a, b) \cap X_1$, but $\mathcal{H}_1 \neq \mathcal{H}_2 | X_1$. If $\mathcal{H}_2 | X_1$ is a proper subset of \mathcal{H}_1 , then T_1 is a proper refinement of $T_2 | X_1$ and it is checked easily that there is a pair of distinct elements $a, b \in X_1$ such that $\text{mrca}_{T_1}(a, b) \neq \text{mrca}_{T_2}(a, b) | X_1$. Therefore, we can assume that there is an element, C_2 say, of $\mathcal{H}_2 | X_1$ that is not an element of \mathcal{H}_1 . Let C_1 be the minimal cluster of T_1 that contains C_2 and let x be an element of $C_1 - C_2$. If x is a label of the vertex of T_1 that corresponds to C_1 , then, by the minimality of C_1 , there is a pair of distinct

elements $a, b \in C_1$ such that $x \in \text{mrca}_{\mathcal{T}_1}(a, b)$, but $x \notin \text{mrca}_{\mathcal{T}_2}(a, b) \mid X_1$. It follows that we can assume also that no element of $C_1 - C_2$ labels the vertex of \mathcal{T}_1 corresponding to C_1 . But then, if c is such a label, it is seen easily that $\text{mrca}_{\mathcal{T}_1}(x, c) \neq \text{mrca}_{\mathcal{T}_2}(x, c) \mid X_1$. This completes the proof of Lemma 3.3.

Lemma 3.4. Let \mathcal{P} be a collection of rooted semi-labeled trees. Let \mathcal{P}' be a set of rooted fully-labeled trees obtained from \mathcal{P} by adding distinct new labels. Then \mathcal{P} is perfectly compatible if and only if \mathcal{P}' is perfectly compatible. Moreover, if \mathcal{T}' is a rooted semi-labeled tree that perfectly displays \mathcal{P}' , then \mathcal{T}' perfectly displays \mathcal{P} .

Proof. Suppose that \mathcal{P} is perfectly compatible and let \mathcal{T} be a rooted semi-labeled tree that perfectly displays \mathcal{P} . For each element $c' \in \mathcal{L}(\mathcal{P}') - \mathcal{L}(\mathcal{P})$, there is a unique rooted fully-labeled tree, \mathcal{T}_1' say, in \mathcal{P}' for which $c' \in \mathcal{L}(\mathcal{T}_1')$. Furthermore, because c' is one of the added labels, c' labels a vertex of \mathcal{T}_1' of degree at least three and so there exist labels a and b of \mathcal{T}_1' such that $\text{mrca}_{\mathcal{T}_1'}(a, b) = \{c'\}$, where \mathcal{T}_1 is the tree in \mathcal{P} corresponding to \mathcal{T}_1' , because all leaves of \mathcal{T}_1 must be labeled.

Now let \mathcal{T}' be the rooted semi-labeled tree obtained from \mathcal{T} by adding the labels of $\mathcal{L}(\mathcal{P}') - \mathcal{L}(\mathcal{P})$ so that if $c' \in \mathcal{L}(\mathcal{P}') - \mathcal{L}(\mathcal{P})$, $c' \in \mathcal{T}_1'$, and $\text{mrca}_{\mathcal{T}_1'}(a, b) = \{c'\}$ for some labels a and b of \mathcal{T}_1 , then $c' \in \text{mrca}_{\mathcal{T}}(a, b)$. By the previous paragraph and the fact that \mathcal{T} perfectly displays \mathcal{P} , it is seen easily that, for all $a, b \in \mathcal{L}(\mathcal{T}_1')$,

$$\text{mrca}_{\mathcal{T}_1'}(a, b) = \text{mrca}_{\mathcal{T}'}(a, b) \mid \mathcal{L}(\mathcal{T}_1')$$

It now follows by Lemma 3.3 that \mathcal{T}' perfectly displays \mathcal{T}_1' and hence \mathcal{P}' .

The rest of the proof of Lemma 3.4 is straightforward and omitted.

Lemma 3.5. Let \mathcal{P} be a collection of rooted fully-labeled trees. If \mathcal{P} is perfectly compatible, then there exists a rooted fully-labeled tree that perfectly displays \mathcal{P} with label set $\mathcal{L}(\mathcal{P})$.

Proof. Suppose that \mathcal{P} is perfectly compatible and let $\mathcal{T} = (T; \phi)$ be a rooted semi-labeled tree that perfectly displays \mathcal{P} and has label set $\mathcal{L}(\mathcal{P})$. Now suppose that among all rooted semi-labeled trees that perfectly display \mathcal{P} and have label set $\mathcal{L}(\mathcal{P})$, the tree T has the least number of unlabeled vertices. If T has no unlabeled vertices, then the lemma is proved. Therefore, assume that there is a vertex, u say, of T that is unlabeled. Because \mathcal{T} is a rooted semi-labeled tree, u has out-degree at least two. Furthermore, because \mathcal{T} perfectly displays \mathcal{P} and \mathcal{P} is a collection of rooted fully-labeled trees, it follows by Lemma 3.3 that there is no tree \mathcal{T}_1 in \mathcal{P} with labels a and b such

that $\text{mrca}_{\mathcal{T}_1}(\phi(a), \phi(b)) = u$. By Lemma 3.3 again, this in turn implies that if v_1, v_2, \dots, v_n are immediate descendants of u in T , then the rooted semi-labeled tree obtained from T by contracting $\{u, v_i\}$ for all i and labeling the identified vertex with $\cup_{i \in \{1, \dots, n\}} \phi_1^{-1}(v_i)$ perfectly displays \mathcal{P} . But the latter tree has one less unlabeled vertex than T . This contradiction completes the proof of the lemma.

It follows from Lemma 3.4 and the description of SEMI-LABELEDBUILD that Theorem 3.1 is an immediate consequence of the following theorem.

Theorem 3.6. *Let \mathcal{P} be a collection of rooted fully-labeled trees. Then FULLY-LABELEDBUILD applied to \mathcal{P} either:*

- (i) *returns a rooted fully-labeled tree that perfectly displays \mathcal{P} if \mathcal{P} is perfectly compatible, or*
- (ii) *returns the statement \mathcal{P} is not perfectly compatible otherwise.*

Proof. First suppose that \mathcal{P} is perfectly compatible. Under this assumption, we show that FULLY-LABELEDBUILD applied to \mathcal{P} outputs a rooted fully-labeled tree. If this is not the case, then FULLY-LABELEDBUILD outputs *not perfectly compatible*, in which case, the associated cluster and root-label graph, G say, has no isolated vertices at some iteration of the algorithm. Let S denote the vertex set of G . Because \mathcal{P} is perfectly compatible, $\mathcal{P}|S$ is perfectly compatible and so, by Lemma 3.5, there exists a rooted fully-labeled tree \mathcal{T} with labeled set S that perfectly displays $\mathcal{P}|S$. Let c be a root label of \mathcal{T} . Then, because \mathcal{T} perfectly displays $\mathcal{P}|S$, every tree of $\mathcal{P}|S$ in which c is a label has the property that c is a root label. Furthermore, because G has no isolated vertices, c must be joined to an element of $S - \{c\}$ in G by some edge and this edge must be a type (iii) edge; for otherwise, c is a non-root label of some tree in $\mathcal{P}|S$. It now follows that there exists a tree \mathcal{T}_1 in $\mathcal{P}|S$ such that a, b, c are distinct vertices of $\mathcal{L}(\mathcal{T}_1)$, $c \in \text{mrca}_{\mathcal{T}_1}(a, b)$, and, in G , there is a path joining a and b . We next show that the existence of this path implies that a and b are elements of a cluster of \mathcal{T} .

Let G_0 denote the graph with vertex set S and the edge set of which consists of all type (i) and (ii) edges of G . Let u and v be any two vertices of G_0 . If $\{u, v\}$ is an edge of this graph, then, using the fact that \mathcal{T} perfectly displays \mathcal{P} , it is checked easily that u and v must be in the same maximal cluster of \mathcal{T} . Clearly, being in the same maximal cluster of a given rooted semi-labeled tree is a transitive relation and so if there is a path in G_0 joining two vertices, then these two vertices are in the same maximal cluster of \mathcal{T} . Now G is obtained from G_0 by adding sets of edges iteratively that join a particular root label of trees in $\mathcal{P}|S$ to all its descendant labels. Let E_1, E_2, \dots, E_k denote the corresponding sequence of these added sets of edges, and

let z_1, z_2, \dots, z_k denote the associated sequence of particular root labels. Let E_0 denote the edge set of G_0 and, for all $i \in \{1, 2, \dots, k\}$, let G_i denote the graph with vertex set S and edge set $E_0 \cup E_1 \cup E_2 \cup \dots \cup E_i$.

Consider the graph G_1 . By the construction of G_1 , there is a rooted fully-labeled tree T_1 in $\mathcal{P}|S$ with root label z_1 and distinct proper descendant labels x_1 and y_1 such that $z_1 \in \text{mrca}_{T_1}(x_1, y_1)$ and, in G_0 , there is a path joining x_1 and y_1 . By the existence of this path and the previous paragraph, x_1 and y_1 are in the same maximal cluster of \mathcal{T} . Because \mathcal{T} perfectly displays $\mathcal{P}|S$ and $z_1 \in \text{mrca}_{T_1}(x_1, y_1)$, it follows that z_1 must also be in this particular maximal cluster of \mathcal{T} . Because all the edges in E_1 contain z_1 and because G_0 has the transitive property mentioned in the last paragraph, G_1 also has the property that if there is a path joining two vertices in G_1 , then these two vertices are in the same maximal cluster of \mathcal{T} . Continuing in this way for G_2, G_3, \dots, G_{k-1} and lastly for G_k , we deduce that G_k , and hence $G(\mathcal{P}|S)$, also have this edge transitive property. But then, because a and b are joined by a path in $G(\mathcal{P}|S)$, a and b are in the same maximal cluster of \mathcal{T} and so $c \notin \text{mrca}_{\mathcal{T}}(a, b)$. This contradiction shows that **FULLY-LABELEDBUILD** does indeed output a rooted fully-labeled tree if \mathcal{P} is compatible.

Now suppose that **FULLY-LABELEDBUILD** outputs a rooted fully-labeled tree \mathcal{T} . Here, we show that \mathcal{T} perfectly displays \mathcal{P} . Let T_1 be a rooted semi-labeled tree in \mathcal{P} with label set X_1 and let $a, b \in X_1$. By Lemma 3.3, it suffices to show that $\text{mrca}_{T_1}(a, b) = \text{mrca}_{\mathcal{T}}(a, b)|X_1$.

We show first that if $c \in \text{mrca}_{T_1}(a, b)$, then $c \in \text{mrca}_{\mathcal{T}}(a, b)|X_1$. Throughout this part of the proof, we freely use the fact that, because $c \in \text{mrca}_{T_1}(a, b)$, we have $c \in \text{mrca}_{T_1}(a, c)$ and $c \in \text{mrca}_{T_1}(b, c)$. Let S be the minimal cluster of \mathcal{T} that contains a, b , and c , and consider the graph $G(\mathcal{P}|S)$. If c is not isolated, then either a and c are in the same cluster of a fully-labeled tree in \mathcal{P} or c is the root label of a fully-labeled tree in \mathcal{P} . In both cases, it follows that a and c are in the same connected component of $G(\mathcal{P}|S)$. Similarly, b and c are in the same connected component of $G(\mathcal{P}|S)$. It now follows that a, b , and c must be in the same connected component of $G(\mathcal{P}|S)$. This implies that S is not the minimal cluster of \mathcal{T} that contains a, b , and c . Therefore, we can assume that c is an isolated vertex of $G(\mathcal{P}|S)$ and, moreover, that it labels the vertex of \mathcal{T} corresponding to S .

If a, b , and c label the same vertex of T_1 , then, by symmetry, a, b , and c are all isolated vertices of $G(\mathcal{P}|S)$ and so a, b , and c label the vertex of \mathcal{T} corresponding to S . Hence, in this case, $c \in \text{mrca}_{\mathcal{T}}(a, b)|X_1$.

Now assume that a and b do not label the same vertex of T_1 , but that b and c do. Then, because c is isolated, it follows by the argument above that b is also an isolated vertex of $G(\mathcal{P}|S)$. Furthermore, b also labels the vertex of \mathcal{T} corresponding to S . Because $c \in \text{mrca}_{T_1}(a, b)$, a is not isolated in $G(\mathcal{P}|S)$ and so $c \in \text{mrca}_{\mathcal{T}}(a, b)|X_1$.

Lastly, assume that no pair of a , b , and c label the same vertex of \mathcal{T}_1 . Because c is isolated, c must have been isolated after Step (ii) in the construction of $G(\mathcal{P} \mid S)$, and so the relation $c \in \text{mrca}_{\mathcal{T}_1}(a, b)$ implies that a and b are joined by a red edge labeled c in Step (iii)(a) of this construction. Moreover, because c remains isolated at the end of the construction, a and b must be in separate connected components of $G(\mathcal{P} \mid S)$. Therefore, $c \in \text{mrca}_{\mathcal{T}}(a, b) \mid X_1$.

We have now established $\text{mrca}_{\mathcal{T}_1}(a, b) \subseteq \text{mrca}_{\mathcal{T}}(a, b) \mid X_1$. It follows from Lemma 3.2 that, for all $a, b \in \mathcal{L}(\mathcal{T}_1)$, a and b label the same vertex of \mathcal{T} precisely if a and b label the same vertex of \mathcal{T}_1 . By the argument in the preceding paragraph, $\text{mrca}_{\mathcal{T}_1}(a, b) \cap \text{mrca}_{\mathcal{T}}(a, b) \mid X_1$ is non-empty. Hence, $\text{mrca}_{\mathcal{T}_1}(a, b) = \text{mrca}_{\mathcal{T}}(a, b) \mid X_1$.

We now consider the running time of SEMI-LABELEDBUILD applied to a collection \mathcal{P} of rooted semi-labeled trees. Because it is more than likely that there is a faster method for determining if \mathcal{P} is perfectly compatible, a detailed analysis is omitted. The point is to show that there exists a polynomial-time algorithm (in the size of $\mathcal{L}(\mathcal{P})$) for determining perfect compatibility of \mathcal{P} .

Let \mathcal{P}' be a collection of rooted fully-labeled trees obtained from \mathcal{P} by adding distinct new labels. Because all vertices of degree at most two are labeled in a rooted semi-labeled tree, the only possible unlabeled vertices are interior vertices with degree at least three. The number of such interior vertices is at most one less than the number of leaves. Summing over all trees in \mathcal{P} , this implies that $|\mathcal{L}(\mathcal{P}')| - |\mathcal{L}(\mathcal{P})| \leq |\mathcal{L}(\mathcal{P})| - 1$. Thus, it suffices to show that the running of SEMI-LABELEDBUILD is polynomial in $|\mathcal{L}(\mathcal{P}')|$. Clearly, the construction of the cluster and root-label graph at each iteration of FULLY-LABELEDBUILD can be done in such a time. Furthermore, because we consider only proper restrictions of the input collection of rooted fully-labeled trees at Step 4 of FULLY-LABELEDBUILD, the number of iterations of FULLY-LABELEDBUILD is bounded by $\mathcal{L}(\mathcal{P}')$. It now follows that the running time of SEMI-LABELEDBUILD is polynomial in the size of $\mathcal{L}(\mathcal{P}')$.

4. Ancestrally displays

If \mathcal{T} is a rooted semi-labeled tree that perfectly displays a collection \mathcal{P} of rooted semi-labeled trees, then \mathcal{T} preserves all the most recent common ancestor relationships described by \mathcal{P} . As a consequence, no polytomies in \mathcal{P} are resolved in \mathcal{T} . Thus, as mentioned in Section 1, the notion of perfectly compatible is very strong. In this section, we introduce a notion of

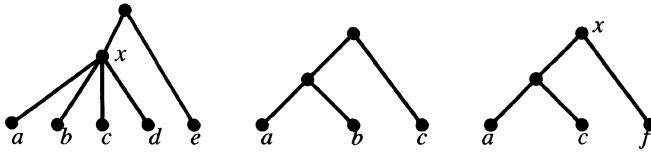


Figure 7. A collection of semi-labeled trees.

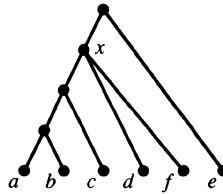


Figure 8. The rooted semi-labeled tree outputted by ANCESTRALBUILD when applied to the rooted semi-labeled trees in Figure 7.

compatibility that allows the resolution of polytomies, but still maintains all the descendancy relationships of a collection of rooted semi-labeled trees. Moreover, we present a polynomial-time algorithm for determining if a collection of rooted semi-labeled trees is compatible under this new notion.

Let $X' \subseteq X$. A rooted semi-labeled tree \mathcal{T} on X *ancestrally displays* a rooted semi-labeled tree \mathcal{T}' on X' if $\mathcal{T} | X'$ refines \mathcal{T}' , and for all $a, b \in X'$, the following hold:

1. if $a <_{\mathcal{T}'} b$, then $a <_{\mathcal{T}} b$, and
2. if a is not comparable to b in \mathcal{T}' under $<_{\mathcal{T}'}$, then a is not comparable to b in \mathcal{T} under $<_{\mathcal{T}}$.

Intuitively, (1) and (2) imply that \mathcal{T} preserves the ancestor-descendant relationships of \mathcal{T}' , but might not preserve the most recent common ancestor relationships of \mathcal{T}' , which is required for the notion of perfectly displays. Consequently, perfectly displays is a stronger notion than ancestrally displays. Each of the rooted semi-labeled trees in Figure 7 is ancestrally displayed by the rooted semi-labeled tree in Figure 8. However, the first tree in Figure 7 is not perfectly displayed by the rooted semi-labeled tree in Figure 8. In comparison with the standard notion of displays, ancestrally displays is stronger because the former does not preserve descendancy. A collection \mathcal{P} of rooted semi-labeled trees is *ancestrally compatible* if there is a rooted semi-labeled tree \mathcal{T} that ancestrally displays every tree in \mathcal{P} , in which case, \mathcal{T} *ancestrally displays* \mathcal{P} .

In this section, we present a polynomial-time algorithm (called ANCESTRALBUILD) for solving the following problem.

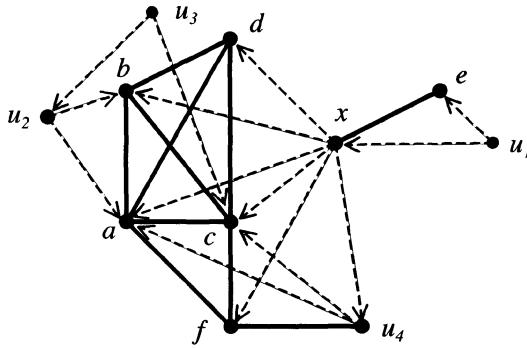


Figure 9. The descendancy graph of \mathcal{P} . Arcs are indicated by dashed lines, edges by solid lines.

Problem: HIGHER TAXA ANCESTOR COMPATIBILITY

Instance: A collection \mathcal{P} of rooted semi-labeled trees.

Question: Does there exist a rooted semi-labeled tree that ancestrally displays \mathcal{P} and, if so, can we construct such a rooted semi-labeled tree?

Before describing ANCESTRALBUILD and its subroutine DESCENDANT, we need first to define a particular graph and a construction. This graph consists of a mixture of arcs (directed edges) and edges. Let \mathcal{P} be a collection of rooted fully-labeled trees. This graph, called the *descendancy graph of \mathcal{P}* and denoted $D(\mathcal{P})$, is defined as follows. The vertex set of $D(\mathcal{P})$ is $L(\mathcal{P})$. The arc set $A(\mathcal{P})$ of $D(\mathcal{P})$ is

$$\{(c, a) : c <_{\mathcal{T}} a \text{ for some } \mathcal{T} \text{ in } \mathcal{P}\},$$

and the edge set $E(\mathcal{P})$ of $D(\mathcal{P})$ is

$$\{\{a, b\} : a \text{ is not comparable to } b \text{ under } \leq_{\mathcal{T}} \text{ for some } \mathcal{T} \text{ in } \mathcal{P}\}.$$

As an example of a descendancy graph, let \mathcal{P} be the collection of fully-labeled trees formed from the trees in Figure 7 by adding u_1 to the root of the leftmost tree, u_3 to the root and u_2 to the other unlabeled vertex of the middle tree, and u_4 to the unlabeled vertex of the rightmost tree. Figure 9 shows the descendancy graph of \mathcal{P} where, to avoid clutter, only edges and arcs from parents to immediate descendants are shown.

The descendancy graph plays an important role in ANCESTRALBUILD. However, unlike SEMI-LABELEDBUILD, where a cluster and root-label graph

is constructed at each iteration, the descendancy graph for \mathcal{P} is constructed just once and then successive iterations consider particular restrictions of it. To this end, we will denote the subgraph of $D(\mathcal{P})$ that is induced by a subset S of the vertex set $\mathcal{L}(\mathcal{P})$ by $D(\mathcal{P})|S$; that is, $D(\mathcal{P})|S$ denotes the subgraph of $D(\mathcal{P})$ obtained by deleting all vertices of $\mathcal{L}(\mathcal{P}) - S$ and their incident arcs and edges. In association with $D(\mathcal{P})$ (or any of its vertex-induced subgraphs), the *in-degree* of a vertex a is the number of arcs directed into a (edges are ignored), and an *arc component* is a connected component of the graph obtained by deleting all edges.

Lastly, we define our construction. Let $\mathcal{T} = (T; \phi)$ be a rooted semi-labeled tree. We say that a rooted semi-labeled tree \mathcal{T}_1 has been obtained from \mathcal{T} by *adding descendants to leaves* if, for each multi-labeled leaf vertex u of T , we adjoin a new leaf vertex v to u by a new edge, and then label each new leaf vertex with a distinct new label. For a collection \mathcal{P} of rooted semi-labeled trees, \mathcal{P}_1 has been obtained from \mathcal{P} by *adding descendants to leaves* if it has been obtained by adding descendants to leaves to every tree in \mathcal{P} so that all the new labels are distinct.

Algorithm: ANCESTRALBUILD(\mathcal{P}, \mathcal{T})

Input: Let \mathcal{P} be a collection of rooted semi-labeled trees.

Output: A rooted semi-labeled tree \mathcal{T} that ancestrally displays \mathcal{P} or the statement \mathcal{P} is not ancestrally compatible.

1. Construct a collection \mathcal{P}' of rooted fully-labeled trees from \mathcal{P} by adding descendants to leaves and then adding distinct new labels to the resulting collection.
2. Construct the descendancy graph $D(\mathcal{P}')$ of \mathcal{P}' .
3. Call the subroutine DESCENDANT($D(\mathcal{P}'), v', \mathcal{T}'$).
4. If DESCENDANT returns no possible labeling, then return \mathcal{P} is not ancestrally compatible.
5. If DESCENDANT returns a rooted semi-labeled tree \mathcal{T}' , then remove the added labels and return the resulting rooted semi-labeled tree \mathcal{T} .

Algorithm: DESCENDANT($D(\mathcal{P}'), v', \mathcal{T}'$)

Input: The descendancy graph of a collection \mathcal{P}' of rooted fully-labeled trees.

Output: A rooted fully-labeled tree \mathcal{T}' with root vertex v' that ancestrally displays \mathcal{P}' or the statement no possible labeling.

1. Let S_0 denote the set of vertices of $D(\mathcal{P}')$ that have in-degree zero and no incident edges.
2. If S_0 is empty, then halt and return *no possible labeling*.
3. Otherwise,
 - (a) Delete the elements of S_0 (and their incident arcs) from $D(\mathcal{P}')$ and denote the resulting graph by $D(\mathcal{P}') \setminus S_0$.
 - (b) Let S_1, S_2, \dots, S_k denote the vertex sets of the arc components of $D(\mathcal{P}') \setminus S_0$.
 - (c) Delete all edges of $D(\mathcal{P}') \setminus S_0$ the end vertices of which are in distinct arc components of this graph.
 - (d) For each element $i \in \{1, 2, \dots, k\}$, call $\text{DESCENDANT}(D(\mathcal{P}') \setminus S_i, T'_i)$. If $\text{DESCENDANT}(D(\mathcal{P}') \setminus S_i, v'_i, T'_i)$ returns a tree, then assign the labels in S_0 to v' and attach T'_i to v' via the edge $\{v'_i, v'\}$.

The general approach of the algorithm DESCENDANT is the same as that of FULLY-LABELEDBUILD. In particular, it attempts to construct a rooted fully-labeled tree that ancestrally displays \mathcal{P}' beginning with the root and moving towards the leaves. To illustrate ANCESTRALBUILD, the rooted semi-labeled tree shown in Figure 8 is the result of applying this algorithm to the collection of rooted semi-labeled trees shown in Figure 7.

Remark.

1. Because DESCENDANT considers proper restrictions of $D(\mathcal{P})$ successively, it is clear that DESCENDANT returns either “no possible labeling” or a rooted semi-labeled tree. ANCESTRALBUILD consequently returns either “ \mathcal{P} is not ancestrally compatible” or a rooted semi-labeled tree.
2. Because every tree in \mathcal{P}' is fully-labeled, it follows that the only labels in S_0 at any iteration are root labels of the corresponding restrictions of \mathcal{P}' . Thus, in regards to the last step of DESCENDANT, $\mathcal{P}' \setminus S_i$ is a rooted fully-labeled tree for all i . This fact will be useful later.

Theorem 4.1. *Let \mathcal{P} be a collection of rooted semi-labeled trees. Then ANCESTRALBUILD applied to \mathcal{P} either:*

- (i) *returns a rooted semi-labeled tree that ancestrally displays \mathcal{P} if \mathcal{P} is ancestrally compatible, or*
- (ii) *returns the statement \mathcal{P} is not ancestrally compatible otherwise.*

The proof Theorem 4.1 makes use of the following lemma. The proof follows the approach used in the proof of Lemma 3.4 and is omitted.

Lemma 4.2. *Let \mathcal{P} be a collection of rooted semi-labeled trees. Let \mathcal{P}' be a set of rooted fully-labeled trees obtained from \mathcal{P} by adding descendants to leaves and then adding distinct new labels to the resulting collection. Then \mathcal{P} is ancestrally compatible if and only if \mathcal{P}' is ancestrally compatible. Moreover, if \mathcal{T}' is a rooted semi-labeled tree that ancestrally displays \mathcal{P}' , then \mathcal{T}' ancestrally displays \mathcal{P} .*

Proof of Theorem 4.1. By Lemma 4.2, it suffices to show that the theorem holds if \mathcal{P} is a collection of fully-labeled trees with no multi-labeled leaf vertices. Suppose that \mathcal{P} is ancestrally compatible, and let \mathcal{T} be a semi-labeled tree that ancestrally displays \mathcal{P} . We show that under this assumption, ANCESTRALBUILD applied to \mathcal{P} outputs a rooted semi-labeled tree. Assume that this is not the case. Then, at some iteration of ANCESTRALBUILD, there is subset S of $\mathcal{L}(\mathcal{P})$ for which all vertices of $D(\mathcal{P})|S$ either have in-degree greater than zero or are incident with an edge. Because \mathcal{T} ancestrally displays \mathcal{P} , it is seen easily that $\mathcal{T}|S$ ancestrally displays $\mathcal{P}|S$. Let P be a path of $\mathcal{T}|S$ from the root to a leaf and consider the first label, y say, that is met on this path. In $D(\mathcal{P})|S$, either y does not have in-degree zero or it is incident with an edge. In the first case, this implies that there is another element, x say, of S such that in some tree of \mathcal{P} we have x is a proper ancestor of y . But y was the first label met in P , and so x is not a proper ancestor of y in $\mathcal{T}|S$ and, in particular, in \mathcal{T} ; a contradiction. Therefore, we can assume that, in $D(\mathcal{P})|S$, y has in-degree zero and is incident with an edge. But then, because $\mathcal{P}|S$ is a collection of rooted fully-labeled trees (see remark above), all trees in $\mathcal{P}|S$ in which y is a label has y as a root label. This means that, in $D(\mathcal{P})|S$, y cannot be incident with any edge. This last contradiction completes this direction of the proof.

For the converse, suppose that ANCESTRALBUILD outputs a rooted semi-labeled tree \mathcal{T} . We show that \mathcal{T} ancestrally displays \mathcal{P} . Let \mathcal{T}_1 be a member of \mathcal{P} , and let a and b be elements of $\mathcal{L}(\mathcal{P})$. If $a <_{\mathcal{T}_1} b$, then, because a is an element of an arc component, there is an arc from a to b in the associated descendancy graph. Because ANCESTRALBUILD returns \mathcal{T} , there must be some iteration at which a is an element of S_0 , but b is a vertex of an arc component of the graph obtained by deleting the elements of S_0 including a . It now follows by the description of the descendancy graph that $a <_{\mathcal{T}} b$.

Next assume that a is not comparable to b in \mathcal{T}_1 . Then, in $D(\mathcal{P})$, the vertices a and b are joined by an edge. Because ANCESTRALBUILD outputs \mathcal{T} , this edge is deleted eventually, but not until a and b are in separate arc components of some restriction of $D(\mathcal{P})$. This implies that, in \mathcal{T} , there is a cluster in which a is an element and not b , and there is a cluster in which b is an element and not a . In other words, a is not comparable to b in \mathcal{T} .

Lastly, let X_1 denote the label set of \mathcal{T}_1 . We complete the converse and thus the proof by showing that $\mathcal{T} | X_1$ refines \mathcal{T}_1 . Let C_1 be a cluster of \mathcal{T}_1 . It suffices to show that C_1 is a cluster of $\mathcal{T} | X_1$. Let X'_1 be the subset of X_1 that labels the vertex u of \mathcal{T}_1 corresponding to C_1 . Because \mathcal{T}_1 is fully-labeled, X'_1 is non-empty. Either X'_1 consists of a single element or u is not a leaf vertex. In the first case, this element is comparable trivially with itself. In the second case, for all $a, b \in X'_1$, there is a label c of \mathcal{T}_1 such that $a <_{\mathcal{T}_1} c$ and $b <_{\mathcal{T}_1} c$, and so, by an earlier argument, $a <_{\mathcal{T}} c$ and $b <_{\mathcal{T}} c$. Hence, a is comparable with b in \mathcal{T} . Furthermore, the same arguments imply that, for all $y \in C_1 - X'_1$, for all $x \in X'_1$, and for all $z \in X_1 - C_1$, we have x is a proper ancestor of y in \mathcal{T} , and either z is a proper ancestor of x or x and z are not comparable in \mathcal{T} . It now follows that C_1 is a cluster of $\mathcal{T} | X_1$.

A very similar analysis to that used to show that the running time of SEMI-LABELEDBUILD is polynomial in the size of $\mathcal{L}(\mathcal{P})$ shows that the running time of ANCESTRALBUILD is also polynomial in the size of $\mathcal{L}(\mathcal{P})$. We leave the details to the reader.

Final Remarks.

1. Some extensions of the problems described in this chapter are considered by Daniel (2004) in his Master's thesis. One in particular is the following. In Figure 6, two of the interior vertices are multi-labeled. For a variety of reasons, such as the labels representing taxa of different levels or the labels representing different taxa of the same rank (e.g., genera), it might have been predetermined that it is not possible for two such labels to label the same vertex. In some cases, such as the former, one way to resolve the problem is to include an additional rooted semi-labeled tree in the input consisting of a root vertex and a leaf where the higher taxon labels the root vertex. However, for many cases, no such resolution is be possible. Hence, a desirable extension to the original problem of HIGHER TAXA COMPATIBILITY is to include a collection of pairs of labels in the instance and then ask the question of whether there exists a rooted semi-labeled tree that perfectly displays \mathcal{P} and has the property for any such pair $\{a, b\}$, a and b label distinct vertices. Surprisingly, Daniel shows that the resulting problem is NP-complete.
2. Both the algorithms described in this chapter are “all-or-nothing” algorithms. Each algorithm returns either a rooted semi-labeled tree with certain properties if one exists or a statement that there is no such tree. In practice, this limits the use of these algorithms. However, we believe that there is a MINCUTSUPERTREE-type approach (see Semple and Steel,

2000; Page, 2002) to resolving this limitation, so that the two algorithms will always output a rooted semi-labeled tree.

Acknowledgements

We thank Olaf Bininda-Emonds, Sebastian Böcker, and Rod Page for their valuable comments. The first author was supported by the New Zealand Institute of Mathematics and its Applications funded program *Phylogenetic Genomics*, and the second author was supported by the New Zealand Marsden Fund.

References

- AHO, A. V., SAGIV, Y., SZYMANSKI, T. G., AND ULLMAN, J. D. 1981. Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. *SIAM Journal on Computing* 10:405–421.
- BRYANT, D., SEMPLE, C., AND STEEL, M. 2004. Supertree methods for ancestral divergence dates and other applications. In O. R. P. Bininda-Emonds (ed). *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 129–150. Kluwer Academic, Dordrecht, the Netherlands.
- DANIEL, P. 2004. *Supertree Methods, Some New Approaches*. M.Sc. thesis, University of Canterbury.
- PAGE, R. D. M. 2002. Modified mincut supertrees. In R. Guigó and D. Gusfield (eds), *Algorithms in Bioinformatics, Second International Workshop, WABI 2002, Rome, Italy, September 17–21, 2002, Proceedings*, pp. 537–552. Springer, Berlin.
- PAGE, R. D. M. 2004. Taxonomy, supertrees, and the Tree of Life. In O. R. P. Bininda-Emonds (ed). *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 247–265. Kluwer Academic, Dordrecht, the Netherlands.
- SEMPLE, C. AND STEEL, M. 2000. A supertree method for rooted trees. *Discrete Applied Mathematics* 105:147–158.
- SEMPLE, C. AND STEEL, M. 2003. *Phylogenetics*. Oxford University Press, Oxford.

Chapter 8

QUARTET SUPERTREES

Raul Piaggio-Talice, J. Gordon Burleigh, and Oliver Eulenstein

Abstract: We introduce two supertree methods that produce unrooted supertrees from unrooted input trees. The methods assemble supertrees from a weighted quartet (four-taxon) tree representation of the input trees. The first method, QLI, extends Willson's local inconsistency quartet method to construct supertrees. This method, which was designed originally to produce a tree from a taxon-character matrix, is not well suited for building accurate supertrees when there is little taxonomic overlap among the input trees. The second method, QILI, builds additionally on Willson's quartet-rectifying process and infers missing phylogenetic information from the input trees. We examined the effectiveness of the quartet-supertree methods using simulated and empirical data sets. These studies suggest that QILI is relatively accurate when compared with the matrix representation with parsimony (MRP) supertree method.

Keywords: phylogeny reconstruction; quartet method; supertree

1. Introduction

Almost all published supertrees (e.g., Bininda-Emonds *et al.*, 1999; Liu *et al.*, 2001; Daubin *et al.*, 2002; Kennedy and Page, 2002; Salamin *et al.*, 2002; Mahon, 2004) have relied on a single supertree construction method, matrix representation with parsimony (MRP; Baum, 1992, Ragan, 1992). We describe two new methods that assemble supertrees from a weighted quartet (four-taxon) tree representation of the input trees. The performance of both methods is compared with MRP in a simulation and an empirical study.

Quartet trees have been used in phylogenetic tree reconstruction for some time (Fitch, 1981; Bandelt and Dress, 1986; Strimmer and von Haeseler,

1996; Ben-Dor *et al.*, 1998; Cao *et al.*, 1998; Willson, 1999; Ranwez and Gascuel, 2001; Robinson-Rechavi and Graur, 2001). Although the optimality criteria and the approaches for combining quartet trees vary among different methods, quartet methods generally combine quartet trees obtained from a character data set. This allows the construction of a quartet tree for all or almost all possible four-taxon subsets of the taxa using optimization criteria such as maximum likelihood or maximum parsimony. By contrast, supertree methods combine input trees that often have little taxonomic overlap. Thus, quartet supertrees would have to be built from sets of quartet trees in which a large number of the possible four-taxon subsets of the joint taxa of the input trees often are not represented. Furthermore, when quartet trees are derived from input trees rather than from a character data set, a new mechanism to weight the quartet trees has to be devised based on their observed frequency in the input trees. We note that RadCon (Thorley and Page, 2000) implements quartet MRP, which uses quartet trees in a variant of the MRP supertree method, but no supertree method has used quartet methods explicitly to build supertrees to our knowledge.

In our quartet-supertree methods, we adopt Willson's (1999) local-inconsistency quartet method. In our first approach, we apply Willson's method directly to the weighted quartet trees obtained from the input trees. We refer to this approach as the Quartet Local Inconsistency (QLI) method. Because of missing quartet trees, however, the supertrees constructed by the QLI method tend to be highly inaccurate. Thus, in a second method, we implement an intermediate step to infer missing quartet trees based on Willson's (2001) rectifying process for quartet trees obtained from character matrices. We refer to this extended approach as the Quartet Inference and Local Inconsistency (QILI) method.

To assess the accuracy of the QLI and QILI methods in contrast to the MRP method, a simulation and an empirical study were performed. The simulation study showed that the QILI method always performed similarly or better than QLI. It also showed that the inference step improved the performance of QILI over QLI greatly when there was lower taxon overlap among the input trees such that the number of missing quartets was higher. Both the empirical and simulation studies showed that the accuracy of the QILI method came close to that of the MRP method. Although the accuracy of the QILI method declined more steeply than that of the MRP method when overlap among the input trees was reduced, the QILI method still performed well when applied to empirical data.

We first provide necessary definitions and notations in Section 2. In Section 3, we survey Willson's approaches and introduce the QLI and QILI supertree methods derived from them. The simulation and empirical studies are described in Section 4. Section 5 presents the results from these studies.

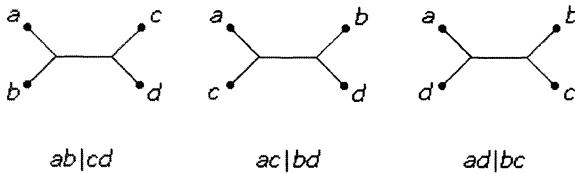


Figure 1. The elements of the set $QT(\{a, b, c, d\})$ representing the three possible quartet trees on the quartet $\{a, b, c, d\}$.

Finally, we discuss the results in Section 6, and review possible future directions for developing quartet-supertree methods further.

2. Definitions

We follow the definitions and notation from Semple and Steel (2003). Unless specified otherwise, all references to a tree throughout the rest of this work will refer implicitly to an unrooted tree.

Let T be a tree. The leaf set of T is $L(T)$, and we say that T is a tree on X if $L(T) = X$. T is a *binary tree* if all non-leaf nodes have degree three. The *restriction* of T to a subset X' of $L(T)$, denoted as $T|X'$, is the minimal subtree T' of T with suppressed degree-two nodes such that $L(T') = X'$. The tree T displays a tree T' if $T|L(T') = T'$.

We say that a set M is a *quartet* if $|M| = 4$ and M is a quartet of the set S if M is a quartet and $M \subseteq S$. A *quartet tree* q is a binary tree where $L(q)$ is a quartet. The notation $ab|cd$ refers to the quartet tree on the set $\{a, b, c, d\}$, in which removing the inner edge yields the trees T_1, T_2 such that $L(T_1) = \{a, b\}$ and $L(T_2) = \{c, d\}$. The *quartet tree set* of a quartet M is $QT(M) = \{q : q$ is a quartet tree on $M\}$, as shown in Figure 1 for the quartet $\{a, b, c, d\}$. The *quartet tree space* of a set S is $Q^*(S) = \bigcup_{M \subseteq S, |M|=4} QT(M)$. The set of *restricted quartet trees* of a tree T is $Q(T) = \{T|M : M$ is a quartet of $L(T)$ and $T|M$ is a quartet tree $\}$. In other words, the set of restricted quartet trees of T is the set of all quartet trees displayed by T . Let Q be a set of quartet trees. Q is *complete* for a set S if Q contains a quartet tree for every quartet of S , and Q is *conflicting* if it contains more than one quartet tree on the same quartet (otherwise it is called *non-conflicting*).

Finally, in the rest of this work, the symbol \mathcal{T} denotes the set of *input trees* that we want to combine into a supertree, and the symbol S denotes the union of their leaf sets.

3. Methods

Our proposed supertree approach decomposes the input trees into the smallest possible relationships among their taxa, and then assembles a single supertree out of these relationships. The smallest relationship that can be obtained among the taxa in the leaf set of a phylogenetic tree requires four elements and can be represented by a quartet tree (Steel, 1992). Therefore, given an input tree set \mathcal{T} , a *quartet-supertree method* analyzes the quartet trees displayed in each element of \mathcal{T} , from which it builds a weighting function for the elements of $Q^*(S)$. This function can then be used as input for a quartet method to produce a tree on S , which is what the QLI method does. Optionally, the weighting function can undergo a preprocessing step to infer missing quartet trees before being passed to the quartet method, which is what the QILI method does. The supertree resulting from either quartet-supertree method is referred to as a *quartet supertree*.

3.1 Quartet decomposition

For a set of \mathcal{T} input trees, our quartet-supertree methods first compute the restricted quartet set $Q(T)$ for each $T \in \mathcal{T}$. Next, the methods count the number of occurrences of each quartet tree $q \in Q^*(S)$ among these sets. If a quartet tree for $L(q)$ is present among the restricted quartet sets, then the frequency $f(q)$ is the number of times q was found divided by the number of times any quartet tree for $L(q)$ was found; otherwise, $f(q) = 0$. Finally, a weight, $w(q)$, is assigned to each quartet tree q as follows:

$$w(q) = \begin{cases} -\log(f(q)) & \text{if } f(q) > 0 \\ r_\infty & \text{if } f(q) = 0 \end{cases},$$

where $r_\infty > \max\{-\log(f(q)) : q \in Q^*(S) \text{ and } f(q) > 0\}$. Note that for every quartet M , such that $M \not\subseteq L(T)$ for any $T \in \mathcal{T}$, the fictitious weight r_∞ is assigned to all three elements of $QT(M)$.

3.2 Tree building

Given a set S and a weighting function w for the elements of $Q^*(S)$, the optimization problem of finding a tree T on S that minimizes $\sum_{q \in Q(T)} w(q)$ follows naturally. This problem is NP-complete because it is a generalization of Maximum Quartet Consistency, which is NP-complete (Berry *et al.*,

1999) even if the given set of quartet trees is non-conflicting and complete for S .

Other alternative heuristic methods have been proposed to build a phylogeny from quartet trees, such as Quartet Puzzling (Strimmer and von Haeseler, 1996; Strimmer *et al.*, 1996) and Willson's (1999) local inconsistency method. The use of a variant of the former as a supertree method has been suggested previously (Pisani and Wilkinson, 2002; Wilkinson *et al.*, 2004), but the latter method is the one that we included in our current supertree methods because it performed similarly or better than Quartet Puzzling in Willson (1999).

3.2.1 Willson's local inconsistency method

Willson's (1999) method takes as input a set S and a weighting function w : $Q^*(S) \rightarrow \mathbf{R}^{\geq 0}$. Given a quartet $M \subseteq S$, we refer to the elements of $Q(M)$ as q_1^M , q_2^M , and q_3^M such that $w(q_1^M) \leq w(q_2^M) \leq w(q_3^M)$. The method starts by choosing the tree q_1^M on a quartet M of S for which $w(q_2^M) - w(q_1^M)$ is maximal. Ties are broken randomly with equal probability. Each iterative step of the method will then take as input a tree T such that $\mathcal{L}(T) \subseteq S$ (initially $|\mathcal{L}(T)| = 4$) and the weighting function w for $Q^*(S)$, and will produce a binary tree T' on $\mathcal{L}(T) \cup \{x\}$ where $x \in S - \mathcal{L}(T)$.

The *excess* of a tree T on a quartet M for a weight function w is defined as:

$$ex_w(T, M) = \begin{cases} w(T|_M) - w(q_2^M) = w(q_1^M) - w(q_2^M) & \text{if } T|_M = q_1^M \\ w(T|_M) - w(q_1^M) & \text{otherwise} \end{cases},$$

and the *local inconsistency* $LI_w(T, x)$ of a tree T on a taxon $x \in \mathcal{L}(T)$ for a weight function w is defined as:

$$LI_w(T, x) = \max \{ex_w(T, M) : M \text{ is a quartet of } \mathcal{L}(T) \text{ and } x \in M\}.$$

At each step, the method considers each taxon $x \in S - \mathcal{L}(T)$, builds all possible binary trees that can be obtained by adding x to T , and stores the trees T_1^x and T_2^x with the maximal and runner-up local inconsistency on x , respectively. The tree that is chosen as T' is the tree T_1^x for the taxon x that maximizes $LI(T_2^x, x) - LI(T_1^x, x)$. This difference indicates how much unambiguous support there is for a particular placement of the chosen taxon. Again, ties are broken randomly with equal probability. When $\mathcal{L}(T') = S$, the method ends with T' as the output; otherwise, T' is passed as input to the next iteration. Note that the random breaking of ties can cause different runs on

the same input to return different supertrees. This behavior is not unique to this approach and also happens in the parsimony heuristic implemented in PAUP* (Swofford, 2002) and used for MRP.

The time complexity of Willson's method can be shown to be $O(n^6)$, where $n = |S|$. Given a set of n taxa, there are $\binom{n}{4} = O(n^4)$ quartets. If we fix one taxon, then there are $\binom{n}{3} = O(n^3)$ quartets. Let m be the number of taxa in the tree T at the beginning of any given step. In that step, there are $n - m$ taxa to try out in $m + 1$ edges of T . For each taxon x and each edge of T , the local inconsistency is computed. This requires checking all the quartets that include taxon x and the other elements of which are members of $\mathcal{L}(T)$. Because the first step of the method begins with a quartet tree and the last step begins with a tree with $n - 1$ taxa, the total running time is

$$O\left(\sum_{m=4}^{n-1} (n-m)(m+1)\binom{m}{3}\right) \leq O\left(\sum_1^n n^2\binom{n}{3}\right) = O(n^6).$$

3.3 Quartet inference

A potential problem when using quartet methods for supertree construction is the difficulty in deducing the best quartet tree for all quartets of S that are not represented in at least one input tree. To assess the best quartet tree for the missing quartets, we present an optional method to infer them from the existing ones. The inference method is built on the rectifying process described in Willson (2001), which was designed originally to strengthen the signal of an input set of non-conflicting and complete quartet trees.

The input to our inference method consists of two sets of quartet trees: a *supported* set and *rejected* set. Both sets and their union must be non-conflicting, but neither of them nor their union needs to be complete. Each element of the supported set represents the quartet tree that we know is the most supported one for its leaf set, and each element of the rejected set represents a quartet tree that is not acceptable for its leaf set (either of the two other quartet trees for that leaf set is considered correct).

To build this kind of input from our weighting function w on $\mathcal{Q}^*(S)$, we proceed to examine the weights of the elements of $QT(M)$ for each quartet M of S . Three outcomes are possible:

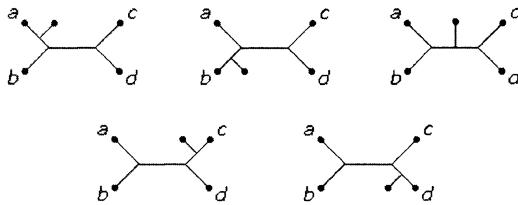


Figure 2. All quintet trees that display the quartet tree $ab \mid cd$ with an additional taxon.

1. $w(q_1^M) < w(q_2^M) \leq w(q_3^M)$. In this case, q_1^M is added to the supported set.
2. $w(q_1^M) = w(q_2^M) < w(q_3^M)$. In this case, q_3^M is added to the rejected set.
3. $w(q_1^M) = w(q_2^M) = w(q_3^M)$. In this case, we have no information to support any of the three quartet trees on M over the other two and so the quartet is ignored.

The inference method consists of assigning to each element of $Q^*(S)$ a *demerit counter*, which is set initially to zero. Then, for each quartet tree $q \in QT(M)$ of each quartet M of S , the method builds all *quintet trees* (binary trees with five leaves) on every *quintet* (subset of size five) of S in which q is displayed (see Figure 2). Note that there are five such trees for each element of $S - M$ (one for each of the edges of q), and that each such quintet tree displays four quartet trees besides q . The demerit counter of q is increased by one for each of these quartet trees that is present in the rejected set and for every different quartet tree for the same quartet that is present in the supported set.

For each quartet M of S , if there is a quartet tree in $QT(M)$ with less demerits than either of the other two, then it is chosen as the correct quartet tree for M and given full support (even if another quartet tree for M was present in the supported input set). Otherwise, if there is a tie between the two quartet trees with the least demerits or if all three quartet trees have the same number of demerits, no decision is made regarding the quartet M . However, the whole process can be repeated taking the output of the previous pass as input for the supported set, and keeping the trees for the quartets that remain unresolved in the rejected set. It is usually the case that demerit ties are resolved eventually.

In our implementation, we repeat the inference method until no further changes are detected or until a maximum number of iterations I is reached. To speed up the computation, we introduce a stability parameter P . If the quartet tree suggested on a given quartet does not change during P consecutive iterations, then it is considered stable and the rest of the

iterations skip that quartet. For all our experiments (see below), we used the values $I = 40$ and $P = 5$.

The output of this inference method is a set of most-supported quartet trees on each quartet. To convert this output into the weighting function for $Q^*(S)$ used in the tree-building method, the fully supported quartet tree for each quartet M is given a weight of 0 and the two other elements of $QT(M)$ are given a weight of r_∞ , where r_∞ is the same as used for the original weighting. For all quartets M for which no quartet tree is supported fully, all three quartet trees on M get a weight of r_∞ . The running time of each iteration of the inference method step is $O(n^5)$, where $n = |S|$. However, the whole process can take a variable amount of time because, depending on the input, the method could converge to a stable set in only one iteration or it could end up being repeated I times.

4. Effectiveness studies

4.1 Simulation study

We first used simulated data sets to compare the accuracy of the QLI and QILI supertree methods relative to the MRP supertree method. We used the simulated data set from Eulensteiner *et al.* (in press; see also Burleigh *et al.*, 2004), to which the reader is referred for full details of the simulation protocol.

Each replicate instance in the data set was simulated starting from a 96-taxon rooted tree (referred to as the *model tree*), which was generated using the default parameters of the YULE_C procedure from the program r8s v1.60 (Sanderson, 2003). An alignment of DNA sequences was derived from the model tree according to the Kimura two-parameter model of evolution (Kimura, 1980) using Seq-Gen v1.2.6 (Rambaut and Grassly, 1997). The alignment was partitioned into 1 000 bp blocks, and each of the 1 000 bp partitions was used to make an input tree. We deleted taxa randomly from each of the input-tree data sets to create different levels of taxonomic overlap among input trees. The simulations used deletion probabilities (d) for each taxon of 25%, 50% and 75%, and sets of 10 and 20 input trees (t). Each resulting alignment partition gave rise to an input tree by using a heuristic search in PAUP* v4.0b10 (Swofford, 2002) with TBR branch swapping, and computing the strict consensus of the equally most parsimonious trees found. We ran 100 replicates for each combination of parameters d and t , and each simulation replicate was based on a different model tree.

Both the quartet and MRP supertree methods used the same input trees; therefore, differences in the performance of each method were not a result of variation in the quality of input trees. QLI and QILI supertrees were generated by unrooting the input tree sets and running our QLI and QILI implementations on the result. MRP supertrees were built by first generating the matrix representation of each input tree and then using PAUP* with the default settings of the heuristic search procedure with TBR branch swapping. Note that the use of rooted input trees in this case could give MRP a slight advantage. To have unrooted results for all three methods, the root was removed from the resulting MRP supertrees.

The *quartet-fit similarity* was used to evaluate the output against the model trees and the sets of input trees. This measure is analogous to the *triplet-fit similarity* from Page (2002; also used in Eulensteine *et al.*, in press), but is based on quartets instead of triplets. For a supertree T_S and a model tree T_M , the quartet-fit similarity from T_S to T_M is defined as:

$$f(T_S, T_M) = 1 - \frac{i + u}{i + s + u},$$

where s is the number of quartets that induce the same quartet tree in both T_S and T_M ; i is the number of quartets that induce different quartet trees in both trees; and u is the number of quartets that induce a quartet tree in T_M , but induce an unresolved star tree in T_S .

The quartet-fit similarity $f(T_S, \mathcal{T})$ from a supertree T_S to the set \mathcal{T} of input trees is the average of the quartet-fit similarity from T_S to each element of \mathcal{T} . In the rest of this work, all values of quartet-fit similarities will be expressed as percentages. Note that the quartet-fit similarity, like the triplet-fit similarity, is not symmetric.

4.2 Empirical study

The characteristics of the input-tree sets from the simulation study seldom matched those of the trees used in real data analysis, where the input trees are of various sizes and have different degrees of overlap. To observe the accuracy of the quartet supertrees with real data, we applied the QILI method to a set of seven input trees from a previously published MRP supertree study of 121 procellariiform seabird taxa (Kennedy and Page, 2002; TreeBASE study and matrix accession numbers S714 and M1139, respectively).

Our study had as input unrooted versions of the same trees that were used in the MRP study (for details, see Kennedy and Page, 2002). The MRP supertrees were found through the heuristic search procedure from PAUP*

using random stepwise addition of taxa on 10 000 replicates, a maximum of 10 000 trees retained, TBR branch swapping, and all other options set to the default settings. Both the MRP heuristic from PAUP* and the QILI method are subject to random decisions and thus can produce different trees in different runs with the same input. Therefore, both methods were run on the input trees ten times each.

Each run of the QILI method produced a single supertree, the quartet-fit similarity of which to the input-tree set was computed. Each run of the MRP heuristic produced 10 000 trees, the average quartet-fit similarity of which to the input-tree set was computed. In addition, the results of each of the MRP searches were summarized as a majority-rule consensus tree of the 10 000 output trees, for which its quartet-fit similarity to the input tree set was computed.

5. Results

5.1 Simulation study

Figure 3 shows the distributions of the quartet-fit similarity for the QILI and MRP supertrees obtained from the simulation data for the $t = 20$ treatment. The results for the $t = 10$ treatment are very similar and are not shown. The distributions for the QLI method are also not shown because QILI performed similarly or better in almost all the cases. For the three methods, Tables 1 and 2 show the average and standard deviation across the 100 runs performed for each set of parameters.

The accuracy of the QLI and QILI methods was similar to the MRP method when the input trees had at least 50% of all taxa (Figure 3; Tables 1 and 2). In the 25% deletion treatment, the QLI and QILI supertrees had average quartet-fit similarities to the model tree and to the input trees that were within one percent of that of the MRP supertrees when there were 20 input trees (Table 2), and within 2.5 percent when there were 10 input trees (Table 1). Our quartet-supertree methods appear to be more sensitive than the MRP method to decreases in the size of the input trees relative to the final supertree (which implies a decrease in taxon overlap). In other words, the scores of MRP supertrees exceeded those of QLI and QILI by an increasing amount as the taxon overlap decreased (Figure 3; Tables 1 and 2). Also, the QILI method was more resistant to decreases in input-tree size than the QLI method: both the QLI and QILI methods performed similarly at the 25% deletion level, but the QILI supertrees had a higher quartet-fit similarity scores than QLI supertrees at 50% and 75% deletion (Tables 1 and 2).

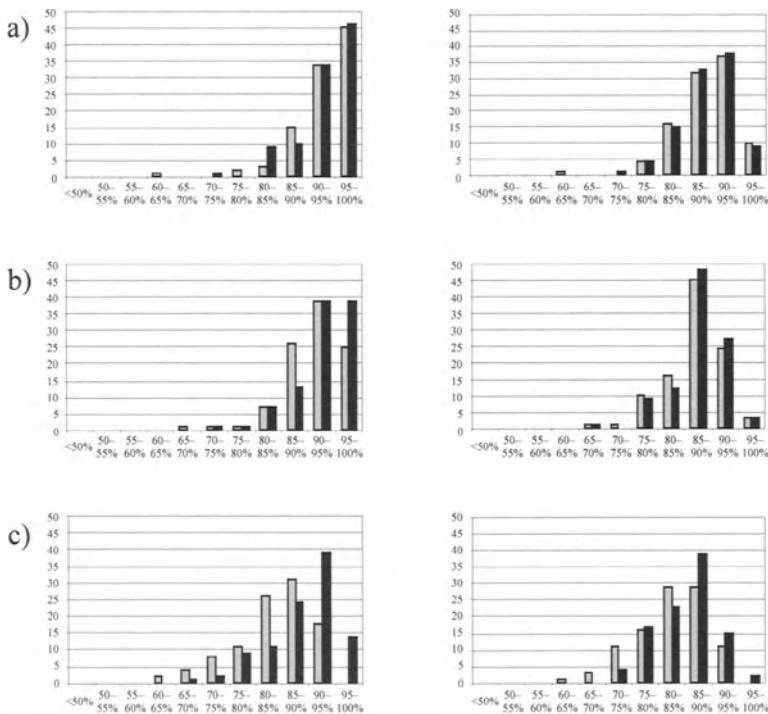


Figure 3. Distributions of the quartet-fit similarity to the model tree (left) and input tree set (right) of QILI (grey) and MRP (black) supertrees. a) With $t = 20$ and $d = 25\%$, b) with $t = 20$ and $d = 50\%$ and c) with $t = 20$ and $d = 75\%$.

When compared with the input trees, the relative accuracy of all the supertree methods was similar with 10 and 20 input trees (Tables 1 and 2, respectively), although they are often slightly higher with 10 input trees. When compared with the model trees, however, the quartet-fit similarity scores of the supertrees were generally higher with more input trees. The difference in quartet-fit similarity scores between the supertrees and the model tree in $t = 10$ (Table 1) and $t = 20$ (Table 2) treatments was most evident when the deletion probabilities were increased.

5.2 Empirical study

The distribution of the quartet-fit similarity scores for the QILI supertrees and the majority-rule consensus of each MRP output are depicted in Figure 4. The average score of each output set of 10 000 MRP supertrees is not shown because it never differs by more than 0.05% from the majority-rule consensus score. In all ten MRP searches, supertrees with the reported

Table 1. Percent quartet-fit similarity to a) input-tree set and b) model tree for various taxon-deletion probabilities with $t = 10$. Avg = average, SD = standard deviation.

a)

	25%		50%		75%	
	Avg	SD	Avg	SD	Avg	SD
QLI	88.92	5.29	85.33	5.68	74.13	7.15
QILI	89.87	4.79	87.40	5.11	83.09	5.72
MRP	89.94	4.56	87.86	4.80	85.16	5.08

b)

	25%		50%		75%	
	Avg	SD	Avg	SD	Avg	SD
QLI	90.00	5.97	85.72	6.37	60.90	9.51
QILI	91.72	5.28	89.30	5.23	74.14	9.10
MRP	92.31	4.94	91.35	4.96	80.45	8.17

Table 2. Percent quartet-fit similarity to a) input-tree set and b) model tree for various taxon-deletion probabilities with $t = 20$. Avg = average, SD = standard deviation.

a)

	25%		50%		75%	
	Avg	SD	Avg	SD	Avg	SD
QLI	88.82	4.92	84.99	6.22	73.64	7.21
QILI	88.77	5.42	86.89	5.27	82.52	6.49
MRP	88.99	4.93	87.29	5.04	84.74	5.51

b)

	25%		50%		75%	
	Avg	SD	Avg	SD	Avg	SD
QLI	92.64	4.51	87.80	7.21	70.33	8.68
QILI	92.78	5.51	91.25	5.51	83.73	6.95
MRP	93.14	5.20	92.92	5.07	88.91	6.37

optimal parsimony length of 213 steps were found (an optimal length of 214 steps for rooted trees was reported by Kennedy and Page, 2002).

The maximum quartet-fit similarity of the QILI supertrees to the input trees was 93.60%, whereas the average was 91.17% with a standard deviation of 2.27%. The best MRP-consensus supertree had a quartet-fit similarity of 95.35% to the input trees, with an average of 94.93% and a standard deviation of 0.29%.

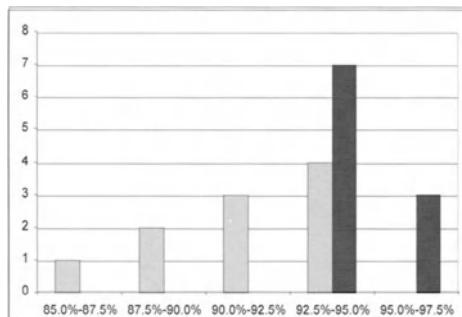


Figure 4. Distribution of the ten quartet-fit similarity scores of QILI supertrees (grey) and MRP majority rule consensus supertrees (black) to the input trees for procellariiform seabirds.

The current implementation of the QILI method used for this study took approximately six hours per run on a standard PC, and the MRP heuristic took approximately half an hour on the same computer.

6. Discussion

6.1 Conclusions

Both the simulation and empirical studies suggest that our quartet-supertree methods, especially QILI, perform well in comparison to the MRP supertree method. The QILI method was able to retain information from the input trees and resolve a supertree with similar accuracy as the MRP method when the input trees contained at least 50% of the total taxa (Tables 1 and 2). It is not entirely surprising perhaps that the performance of the two quartet methods presented dropped off considerably when the input trees contained only 25% of the original taxa (75% deletion) because, in this case, many of the quartets in the unified taxa set were not observed in any input tree. In these cases, the quartet-inference step improved the results noticeably, but the most supported relationship among the elements of these quartets still might not always be estimated accurately.

The design of the simulation study might overemphasize the sensitivity of quartet-supertree methods to taxon sampling. However, the empirical study might be a better example of the circumstances under which supertrees are usually computed. In real supertree studies, the input trees have different sizes, often including some very large ones. For example, six of the seven input trees in the seabird supertree study contained between 15 and 33 (or

12% and 27%, respectively) of the total taxa, but one input tree contained 90 (74%) of the total taxa (Kennedy and Page, 2002). In this case, although most input trees are small, all the quartets combining any of the 90 taxa in the largest input tree will have an informative weight on their quartet trees. Consequently, the average quartet-fit similarity of the QILI supertrees to the input trees in the seabird study (91.17%) is greater than the average scores to the input trees in all the simulation experiments (Tables 1a and 2a), and is most comparable to the 25% deletion treatment in the simulation study. Therefore, the relative size of the largest input tree is probably more indicative of the performance of the quartet-supertree method than the average relative size of the input trees. The benefits of including at least one large input tree are not exclusive to quartet-supertree methods (Bininda-Emonds and Sanderson, 2001), but because missing quartets appear to be problematic for quartet-supertree methods, including such an input tree might be especially crucial for building quartet supertrees. We suggest that one should interpret the placement of taxa that are only present in small input trees with caution in quartet supertrees. Still, with the rapid increase in available sequence data, new methods have been developed for identifying large concatenated data sets from sequence databases (Sanderson *et al.*, 2003) and for inferring large phylogenies (e.g., Huelsenbeck *et al.*, 2001). Thus, we expect that large input trees will be increasingly common in supertree studies.

6.2 Future directions

The QLI and QILI methods constitute a first implementation of quartet-supertree methods. There are several possible enhancements to the current implementations that could improve their speed and accuracy.

First, because the input to supertree methods is usually a set of rooted trees, a rooted tree is often expected as output. In the QLI and QILI methods, the rooting information is lost because these methods work with (unrooted) quartet decomposition. The extensive literature on quartet methods prompted us to choose quartets as a way to assess initially the validity of supertree methods based on the decomposition of the input trees into minimal subtrees. Having established experimentally that this approach is valid, the development of analogous methods that work with rooted trees and a supporting theoretical framework is now a priority. Such methods would decompose the input trees into triplets instead of quartets, which would also imply automatically a linear increase in running time because, given a set of n taxa, there are $\binom{n}{4} = O(n^4)$ quartets, but only $\binom{n}{3} = O(n^3)$ triplets.

A further improvement in speed might come by modifying the methods so that they are able to build a supertree by processing only the quartets (or triplets) present in the input trees without the need for inferring accurate relationships for all possible quartets (or triplets) of the unified taxa set. This method would handle a reduced amount of information and therefore would probably run faster than the current implementations of QLI and QILI.

The accuracy of the resulting supertree could be improved by adding a step similar to the hill-climbing search via branch-swapping operations used in PAUP* and used currently in the heuristic search procedure for MRP. Nonetheless, it is encouraging that the current implementations of the QILI method often performed similarly to the MRP method even without a hill-climbing component. Additionally, in contrast to the MRP method, by using Willson's local inconsistency method, we ensure that the outputs of both the QLI and QILI methods consist of only one tree (see Ross and Rodrigo, 2004). However, like the MRP heuristic, any random decisions in the case of ties means that different runs on the same input can produce different trees. Several runs of a quartet-supertree method might therefore be necessary to ensure that a supertree is found that is close to the globally optimal one. All the supertrees obtained could also be combined via consensus, as is often done with the output trees of other supertree methods that do not return single trees.

As an alternative, other quartet tree building methods could be used instead of the local inconsistency method. As mentioned in Pisani and Wilkinson (2002) and Wilkinson *et al.* (2004), a candidate method is the Quartet Puzzling heuristic (Strimmer and von Haeseler, 1996; Strimmer *et al.*, 1996), which has not been tested on incomplete quartet sets to the best of our knowledge. Other existing quartet methods should also be considered (e.g., Ben-Dor *et al.*, 1998; Bryant and Steel, 2001). Furthermore, alternative methods such as quartet cleaning (Jiang *et al.*, 1998; Berry *et al.*, 1999, 2000; Della Vedova and Wareham, 2002) or dyadic inference (Erdős *et al.*, 1999) could be used also as a basis to infer missing quartet trees.

Quartet-supertree methods could also consider weighting processes that incorporate confidence measures of each input tree or quartet tree. Weighting different clades in input trees based on confidence measures increases the accuracy of MRP methods in simulation (Bininda-Emonds and Sanderson, 2001). Representing whole input-tree confidence values can be achieved in the current setting by considering the input as a multiset of trees (instead of a set), and including each tree a number of times proportional to its confidence value. Observing quartet-tree weights, however, would require small modifications to the methods, which assume implicitly a 100% confidence in each quartet tree displayed in the input trees currently. It is

simple to weight each quartet tree based on either bootstrap support or the Bayesian posterior probability of each quartet tree. This could be done by decomposing the set of bootstrap trees or the trees sampled from a Bayesian Markov chain Monte Carlo simulation. It is also possible to incorporate either prior probabilities or constraints into the weighting procedure for quartet trees. Constraints could help to resolve many of the quartets that are not represented in any input trees, and thus might improve the accuracy of quartet-supertree methods when the input trees are small.

Finally, the input of a quartet-supertree method could be mixed, consisting of character information and agreed-upon skeleton trees the taxon sets of which overlap with the ones in the character set. Quartet trees can be derived from the character set using maximum likelihood or maximum parsimony methods (as is usually the case when applying quartet methods) and from the skeleton trees by unrooting them and decomposing them. Such a mixed approach would be especially useful if the skeleton trees were built from sources other than character data. Its output would be a supertree of the skeleton trees with additional taxa present in the character set, which could indicate a final supertree of the unified taxa set or could help detect errors in the skeleton trees.

7. Software availability

Quartet Suite is a software package that implements the QLI and QILI methods and can be downloaded from our supertree server at <http://genome.cs.iastate.edu>. Quartet Suite's distribution includes binary files for DOS / Windows and for Mac OS X, as well as the source code and usage documentation. The current implementation reads basic NEXUS format files (Maddison *et al.*, 1997).

Acknowledgements

We thank Stephen Willson for his remarks and discussion on quartet methods; Olaf Bininda-Emonds, David Fernández-Baca, Gavin Naylor, and Mark Wilkinson for providing comments on the manuscript; Mike Sanderson for his insight in supertree methods and help in the design of the effectiveness studies; Duhong Chen for his assistance with the simulation study; and Amy Driskell for her assistance with the empirical study.

References

- BANDELT, H.-J. AND DRESS, A. W. M. 1986. Reconstructing the shape of a tree from observed dissimilarity data. *Advances in Applied Mathematics* 7:309–343.
- BAUM, B. R. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon* 41:3–10.
- BEN-DOR, A., CHOR, B., GRAUR, D., OPHIR, R., AND PELLEG, D. 1998. Constructing phylogenies from quartets: elucidation of eutherian superordinal relationships. *Journal of Computational Biology* 5:377–390.
- BERRY, V., JIANG, T., KEARNEY, P., LI, M., AND WAREHAM, H. T. 1999. Quartet Cleaning: improved algorithms and simulations. In J. Nešetřil (ed.), *Algorithms — ESA'99: 7th Annual European Symposium, Prague, Czech Republic, July 1999*, Lecture Notes in Computer Science 1643:313–324. Springer-Verlag, Berlin.
- BERRY, V., BRYANT, D., JIANG, T., KEARNEY, P., LI, M., WAREHAM, H. T., AND ZHANG, H. 2000. A practical algorithm for recovering the best supported edges of an evolutionary tree. In D. Shmoys (ed.), *Symposium on Discrete Algorithms. Proceedings of the Eleventh Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 287–296. Society for Industrial and Applied Mathematics, Philadelphia, PA.
- BININDA-EMONDS, O. R. P., GITTLEMAN, J. L., AND PURVIS, A. 1999. Building large trees by combining phylogenetic information: a complete phylogeny of the extant Carnivora (Mammalia). *Biological Reviews* 74:143–175.
- BININDA-EMONDS, O. R. P. AND SANDERSON, M. J. 2001. An assessment of the accuracy of MRP supertree construction. *Systematic Biology* 50:565–579.
- BRYANT, D. AND STEEL, M. A. 2001. Constructing optimal trees from quartets. *Journal of Algorithms* 38:237–259.
- BURLEIGH, J. G., EULENSTEIN, O., FERNÁNDEZ-BACA, D., AND SANDERSON, M. J. 2004. MRF supertrees. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 65–85. Kluwer Academic, Dordrecht, the Netherlands.
- CAO, Y., ADACHI, J., AND HASEGAWA, M. 1998. Comment on the quartet puzzling method for finding maximum-likelihood tree topologies. *Molecular Biology and Evolution* 15:87–89.
- DAUBIN, V., GOUY, M., AND PERRIERE, G. 2002. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Research* 12:1080–1090.
- DELLA VEDOVA, G. AND WAREHAM, H. T. 2002. Optimal algorithms for local vertex quartet cleaning. *Bioinformatics* 18:1297–1304.
- ERDŐS, P. L., STEEL, M. A., SZÉKELY, L. A., AND WARNOW, T. J. 1999. A few logs suffice to build (almost) all trees (Part 1). *Random Structures and Algorithms* 14:153–184.
- EULENSTEIN, O., CHEN, D., BURLEIGH, J. G., FERNÁNDEZ-BACA, D., AND SANDERSON, M. J. In press. Performance of flip-supertrees. *Systematic Biology*.
- FITCH, W. M. 1981. A non-sequential method of constructing trees and hierarchical classifications. *Journal of Molecular Evolution* 18:30–37.
- HUELSENBECK, J. P., RONQUIST, F., NIELSEN, R., AND BOLLBACK, J. P. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 294:2310–2314.
- JIANG, T., KEARNEY, P., AND LI, M. 1998. Orchestrating quartets: approximation and data correction. In *Proceedings, 39th Annual Symposium on Foundations of Computer Science: November 8–11, 1998, Palo Alto, California*, pp. 416–425. IEEE Computer Society Press, Los Alamitos, California.

- KENNEDY, M. AND PAGE, R. D. M. 2002. Seabird supertrees: combining partial estimates of procellariiform phylogeny. *The Auk* 119:88–108.
- KIMURA, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* 16:111–120.
- LIU, F.-G. R., MIYAMOTO, M. M., FREIRE, N. P., ONG, P. Q., TENNANT, M. R., YOUNG, T. S., AND GUGEL, K. F. 2001. Molecular and morphological supertrees for eutherian (placental) mammals. *Science* 291:1786–1789.
- MADDISON, D. R., SWOFFORD, D. L., AND MADDISON, W. P. 1997. NEXUS: an extensible file format for systematic information. *Systematic Biology* 46:590–621.
- MAHON, A. S. 2004. A molecular supertree of the Artiodactyla. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 411–437. Kluwer Academic, Dordrecht, the Netherlands.
- PAGE, R. D. M. 2002. Modified mincut supertrees. In R. Guigó and D. Gusfield (eds), *Algorithms in Bioinformatics, Second International Workshop, WABI 2002, Rome, Italy, September 17–21, 2002, Proceedings*, pp. 537–552. Springer, Berlin.
- PISANI, D. AND WILKINSON, M. 2002. Matrix representation with parsimony, taxonomic congruence, and total evidence. *Systematic Biology* 51:151–155.
- RAGAN, M. A. 1992. Phylogenetic inference based on matrix representation of trees. *Molecular Phylogenetics and Evolution* 1:53–58.
- RAMBAUT, A. AND GRASSLY, N. C. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer Applications in the Biosciences* 13:235–238.
- RANWEZ, V. AND GASCUEL, O. 2001. Quartet-based phylogenetic inference: improvements and limits. *Molecular Biology and Evolution* 18:1103–1116.
- ROBINSON-RECHAVI, M. AND GRAUR, D. 2001. Usage optimization of unevenly sampled data through the combination of quartet trees: an eutherian draft phylogeny based on 640 nuclear and mitochondrial proteins. *Israel Journal of Zoology* 47:259–270.
- ROSS, H. A. AND RODRIGO, A. G. 2004. An assessment of matrix representation with compatibility in supertree construction. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 35–63. Kluwer Academic, Dordrecht, the Netherlands.
- SALAMIN, N., HODKINSON, T. R., AND SAVOLAINEN, V. 2002. Building supertrees: an empirical assessment using the grass family (Poaceae). *Systematic Biology* 51:136–150.
- SANDERSON, M. J. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* 19:301–302.
- SANDERSON, M. J., DRISKELL, A. C., REE, R. H., EULENSTEIN, O., AND LANGLEY, S. 2003. Obtaining maximal concatenated data sets from large sequence databases. *Molecular Biology and Evolution* 20:1036–1042.
- SEMPLE, C., AND STEEL, M. A. 2003. *Phylogenetics*. Oxford University Press, Oxford.
- STEEL, M. A. 1992. The complexity of reconstructing trees from qualitative characters and subtrees. *Journal of Classification* 9:91–116.
- STRIMMER, K., GOLDMAN, N., AND VON HAESELER, A. 1996. Bayesian probabilities and quartet puzzling. *Molecular Biology and Evolution* 14:210–211.
- STRIMMER, K. AND VON HAESELER, A. 1996. Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. *Molecular Biology and Evolution* 13:964–969.
- SWOFFORD, D. L. 2002. *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Version 4. Sinauer, Sunderland, Massachusetts.

- THORLEY, J. L. and PAGE, R. D. M. 2000. RadCon: phylogenetic tree comparison and consensus. *Bioinformatics* 16:486–487.
- WILKINSON, M., THORLEY, J. L., PISANI, D., LAPOINTE, F.J., AND MCINERNEY, J. O. 2004. Some desiderata for liberal supertrees. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 227–246. Kluwer Academic, Dordrecht, the Netherlands.
- WILLSON, S. J. 1999. Building phylogenetic trees from quartets by using local inconsistency measures. *Molecular Biology and Evolution* 16:685–693.
- WILLSON, S. J. 2001. An error correcting map for quartets can improve the signals for phylogenetic trees. *Molecular Biology and Evolution* 18:344–351.

Chapter 9

BAYESIAN SUPERTREES

Fredrik Ronquist, John P. Huelsenbeck, and Tom Britton

Abstract: In this chapter, we develop a Bayesian approach to supertree construction. Bayesian inference requires that prior knowledge be specified in terms of a probability distribution and incorporates this evidence in new analyses. This provides a natural framework for the accumulation of phylogenetic evidence, but it requires that phylogenetic results be expressed as probability distributions on trees. Because there are so many possible trees, it is usually not feasible to estimate the probability of each individual tree. Therefore, Bayesians summarize the distribution typically in terms of taxon-bipartition frequencies instead. However, bipartition frequencies are related only indirectly to tree probabilities. We discuss two ways in which taxon-bipartition frequencies can be translated into sets of multiplicative factors that function as keys to the probability distribution on trees. The Weighted Independent Binary (WIB) method associates factors to the presence or absence of taxon bipartitions, whereas the Weighted Additive Binary (WAB) method has factors with graded responses dependent on the degree of conflict between the tree and the partition. Although the methods are similar, we found that WAB is superior to WIB. We discuss several ways of estimating WAB factors from partition frequencies or directly from the data. One of these methods suggests a similarity between WAB factors and the decay index; indeed, the WAB factors represent a more natural measure of clade support than the bipartition frequencies themselves or the decay index and its probabilistic analog. WAB factors provide an efficient and convenient way of retrieving prior tree probabilities and WAB supermatrices accurately describe fully statistically specified supertree spaces that can be sampled using MCMC algorithms with the computational efficiency of parsimony. This should allow construction of Bayesian supertrees with thousands of taxa.

Keywords: Bayesian inference; Markov chain Monte Carlo; phylogeny; supertree; tree probability distribution; WAB factor

1. Introduction

Supertree construction is the meta-analysis of phylogenetics: results from the analyses of several smaller data sets are pieced together into larger hypotheses of relationships (Sanderson *et al.*, 1998). It differs from ordinary phylogenetic analysis of large composite matrices in that it combines the results of smaller analyses instead of combining the underlying data. Supertree construction can be used to build very large trees from partially overlapping analyses, and it can also be applied in some situations when ordinary methods cannot. For instance, supertrees might be the only way of combining results from incompatible approaches, such as statistical analyses of discrete data and distance analyses of DNA-DNA hybridization data.

Supertree methods can be fast polynomial algorithms for amalgamating trees (e.g., Semple and Steel, 2000; Steel *et al.*, 2000; Page, 2002). More commonly, however, they are based on parsimony analysis of simplified matrices designed to describe phylogenetic trees or phylogenetic results so that they can be used as building blocks in constructing larger synthetic trees (Baum, 1992; Ragan, 1992). This method is referred to usually as matrix representation with parsimony analysis (MRP). Several refined methods exist for deriving appropriate matrix representations (e.g., Purvis, 1995; Ronquist, 1996; Bininda-Emonds and Bryant, 1998; Bininda-Emonds and Sanderson, 2001; see Baum and Ragan, 2004). In this chapter, we focus on the recently introduced Bayesian approach to phylogenetics and the new perspective it offers on supertree construction in general and on matrix or factorial representation of phylogenetic results in particular.

2. Bayesian inference

Bayesian inference is based on Bayes's rule, which can be formulated

$$(1) \quad f(\theta | X) = \frac{f(\theta) f(X | \theta)}{f(X)},$$

where X represents the observations and θ is a vector of model parameters. In a phylogenetic problem, θ would include the topology of the tree, τ , as well as other parameters, and X would be the data matrix. The function $f(\theta)$ is known as the *prior-probability distribution*, or simply the *prior*, and specifies the probability of the parameter values before the present data were collected. The function $f(X | \theta)$ is referred to as the *likelihood function*; it is this function that is maximized in maximum likelihood (ML) inference. The

denominator $f(X)$, which is a normalizing constant, is the marginal (or total) probability of the data, and is obtained by integrating and summing over all model parameters. Finally, the function $f(\theta|X)$ is the *posterior-probability distribution*, or simply the *posterior*, which describes the probability of the model parameters after the observed data have been taken into account. In Bayesian inference, all conclusions are based on the posterior-probability distribution.

Typically, it is not possible to calculate the posterior analytically. Instead, it is sampled using Markov chain Monte Carlo (MCMC) methods. It is beyond the scope of the current chapter to provide more detailed coverage of the Bayesian MCMC approach in general and its application to phylogenetic inference in particular. The interested reader is referred to recent papers for more detailed introductions and references to pertinent literature (Lewis, 2000; Huelsenbeck *et al.*, 2001, 2002).

3. Bayesian accumulation of evidence

In the context of constructing supertrees, a particularly intriguing aspect of Bayesian inference is that it provides a coherent framework for accumulating scientific knowledge. Because the start and end points of a Bayesian analysis are both formal probability distributions, we can easily use the results of one analysis as the starting point for another. Indeed, it can be shown that such a stepwise approach is equivalent to a simultaneous Bayesian analysis of all the available data given that the initial prior is the same. To see this, assume that we have two sets of data, X_A and X_B , and start with the prior $f(\theta)$. In the analysis of the first set of data, X_A , we obtain the posterior

$$(2) \quad f(\theta|X_A) = \frac{f(\theta)f(X_A|\theta)}{f(X_A)}.$$

If we use this posterior as the prior in the analysis of the second data set, X_B , we obtain the final posterior

$$(3) \quad f(\theta|X_B, X_A) = \frac{f(\theta|X_A)f(X_B|\theta)}{f(X_B)}.$$

By expanding $f(\theta|X_A)$ according to equation (2), we obtain (given that X_A and X_B are independent)

$$(4) \quad f(\theta | X_B, X_A) = \frac{f(\theta) f(X_A | \theta)}{f(X_A)} \frac{f(X_B | \theta)}{f(X_B)} = \frac{f(\theta) f(X_A, X_B | \theta)}{f(X_A, X_B)}.$$

Thus, the final posterior-probability distribution is the same as the one we would have obtained in a single Bayesian analysis of both data sets combined. Note how the final posterior in the combined analysis is obtained by multiplying the likelihood functions $f(X_A | \theta)$ and $f(X_B | \theta)$ with the prior after appropriate normalization.

Now, consider how knowledge about topology is accumulated in the framework of Bayesian phylogenetic inference (Figure 1). Before the analysis is started, a prior-probability distribution on trees is formulated. In the absence of background knowledge, all possible trees are given the same prior probability typically, a so-called uniform prior (Figure 1a). In most cases, different trees have very different probabilities of generating the observed data. These probabilities, or likelihoods, are simply multiplied with the corresponding prior probabilities to obtain the posterior probability of each tree. If the probabilities are measured on the log scale, we simply add the prior and likelihood scores to obtain the posterior score (Figure 1b). This is not a true probability distribution, however, because the probabilities do not sum to 1. After rescaling the values so that they do sum to 1, this is our posterior-probability distribution on trees (Figure 1c). Note that the normalization affects just the position of the baseline when probabilities are measured on the log scale; it leaves the relative tree probabilities (i.e., the absolute difference between log probabilities) unaffected. This is because normalization involves multiplication by a scaling factor, which is equivalent to addition of a constant value on the log scale. Now, if this posterior (Figure 1c) is used as the prior in a subsequent analysis, we need to add only the log-likelihood scores of the trees under the new data on top of the previous (unnormalized or normalized) posterior scores (Figure 1d). Renormalization results in the final posterior-probability distribution (Figure 1e). This distribution is equivalent to the posterior that would have been obtained in simultaneous Bayesian analysis of the two data sets combined (Figure 1f).

The above description is somewhat simplified because topology does not specify a stochastic phylogenetic model fully. Typically, the parameter vector θ of such a model also includes branch lengths and substitution model parameters. We will return to this complication towards the end of the chapter; for now, we will ignore the other parameters and assume that θ contains only the topology (i.e., $\theta = \tau$).

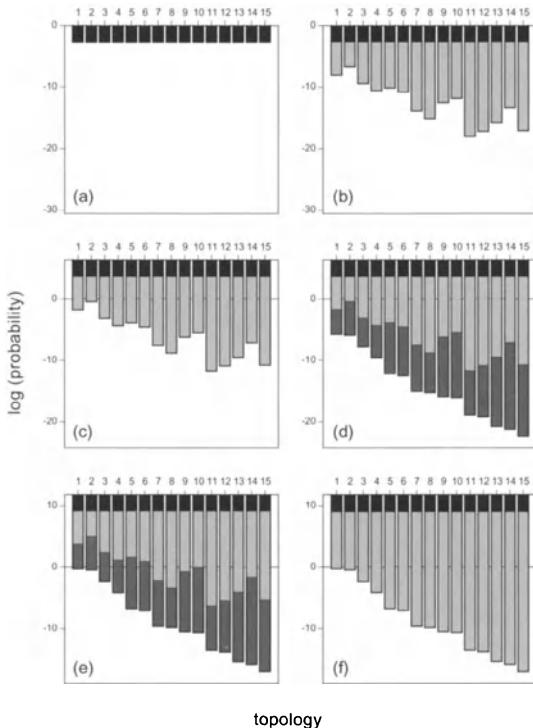


Figure 1. Accumulation of evidence in Bayesian phylogenetic inference. All probabilities are given on a log scale and, because all are <1.0 , the log probabilities are negative. a) Before phylogenetic analysis, a prior-probability distribution on trees is specified. In the lack of background information, all possible trees are associated typically with the same probability (a so-called diffuse prior). b) Different trees usually have very different probabilities (gray bars) of generating the observed data. These probabilities are multiplied with the prior probabilities to give the unnormalized posterior-probabilities of each tree. This is equivalent to adding these quantities on the log scale. c) Normalization (i.e., dividing all tree probabilities with a scaling factor so that they sum to 1.0) yields the posterior-probability distribution on trees. On the log scale, division by a constant scaling factor translates to shifting the position of the baseline. d) The posterior-probability distribution can be used now as the prior in a subsequent analysis. The new posterior scores are obtained by multiplying the prior of each tree (the old posterior) with the probability that the tree generated the new data. On the log scale, the new tree probabilities are added simply on top of the old posterior. e) Normalization again yields our final posterior-probability distribution on trees. f) Exactly the same final posterior distribution would have been obtained if both data sets had been used in a simultaneous analysis given that the initial prior was the same.

Returning to the Bayesian accumulation of phylogenetic evidence (Figure 1), we have seen that it revolves around distributions of tree probabilities. To use the results of a Bayesian phylogenetic analysis of a data set X_i as the

prior in a new analysis, we need to know the posterior-probability distribution on trees, $f(\tau | X_i)$. Building a Bayesian supertree is similar: we need to know the likelihood scores, $f(X_i | \tau)$, of the trees in each analysis. The individual tree scores are combined with an appropriate no-data prior to obtain the (unnormalized) posterior-probability distribution on supertrees. Thus, Bayesian-supertree construction would be straightforward if we could estimate individual tree scores efficiently. Unfortunately, this is usually not feasible computationally.

The MCMC technique allows us to obtain samples from the posterior-probability distribution. After the chain has been run sufficiently long, the proportion of MCMC samples containing a particular tree is a valid estimate of the relative posterior probability falling on that tree. If there are large numbers of trees, however, it takes many MCMC samples before all trees are represented, so the sheer size of tree space usually prohibits all attempts to obtain the complete probability distribution on trees. An additional complication is that the distribution is nearly always skewed strongly, with a tiny fraction of the possible trees comprising most of the probability density. For instance, it is not uncommon to find that 90% or more of the MCMC samples from a small data set consist of only two or three distinct trees, even if there are thousands or millions of trees that are potentially possible. It is only the optimal or near-optimal trees for which we can hope to estimate the posterior probability accurately using standard MCMC sampling. In large analyses with hundreds of taxa, there might even be too many near-optimal trees to estimate the relative probability of any one of them with acceptable accuracy.

In practice, the usual way of obtaining a picture of the posterior-probability distribution on trees is to focus on the frequency of taxon bipartitions instead of on individual tree probabilities. It is convenient to think of the taxon-bipartition frequencies as the posterior probabilities of the corresponding clades although, strictly speaking, the partitions do not represent true independent parameters. However, the partition frequencies do have a formal interpretation.

Let s_i be the frequency of (or support for) a partition i and let b_i be a vector of zeros and ones defining how the partition separates the n taxa into two sets, with the first taxon being in set 0 by convention. For instance $b_1 = \{0, 0, 1, 1\}$ is a partition dividing four taxa into two sets, one with taxa 1 and 2 and the other with taxa 3 and 4. Let $T(b_i^+)$ be the set of all fully bifurcating topologies τ for n taxa consistent with taxon bipartition b_i and let $T(b_i^-)$ be the set of all topologies inconsistent with the bipartition. These are mutually exclusive sets; that is, $T(b_i^+) \cap T(b_i^-) = \emptyset$. Now, we can define the ratio between the support for and against a bipartition in the following *partition-odds* formulation:

$$(5) \quad o_i = \frac{s_i}{1-s_i} = \frac{\sum_{\tau \in T(b_i^+)} f(X|\tau)}{\sum_{\tau \in T(b_i^-)} f(X|\tau)}.$$

Provided that neither s_i nor $1 - s_i$ is close to zero, it is possible to estimate the partition odds, and the ratio between them, accurately using MCMC sampling. Even so, each bipartition frequency is only a statement about the probability ratio between two mutually exclusive classes of trees. To use the results of one Bayesian analysis as a prior in a subsequent analysis or as raw data in constructing a Bayesian supertree, we need to know the individual tree probabilities. A central problem then is how to obtain approximate individual tree probabilities efficiently and accurately from bipartition frequencies, or, in other words, how to use the bipartition frequencies best as a key to the entire probability distribution on trees.

4. Weighted independent binary representation

A simple solution to the above problem is to associate each scored partition b_i with an independent factor $r_i > 1$ specifying how much more probable a tree with the partition is compared with a tree without it (Huelsenbeck *et al.*, 2002). For reasons that will become obvious later on, we refer to this as the *Weighted Independent Binary* (WIB) representation method. A single WIB factor r divides tree space into two classes of trees: those with the corresponding taxon bipartition b and those without it. A tree in the first class is r times more probable than a tree in the second class. If, as we will assume, the initial prior puts equal probability on all trees, then this statement is true for both the relative posterior probabilities $f(\tau|X)$ and the relative likelihood scores $f(X|\tau)$ of the trees. To calculate a factor r from the corresponding bipartition frequency s , assuming that the latter is the only factor changing the prior-probability distribution on trees, we need only know the number of trees in each class: $|T(b_i^+)|$ and $|T(b_i^-)|$. We have

$$(6) \quad \frac{s}{1-s} = \frac{r |T(b_i^+)|}{|T(b_i^-)|}$$

and hence

$$(7) \quad r = \frac{\left| T(b_i^-) \right|}{\left| T(b_i^+) \right|} \frac{s}{(1-s)}.$$

The number of fully bifurcating trees supporting and opposing the taxon bipartition can be calculated easily. Suppose that $B(n)$ is the number of fully bifurcating, unrooted trees for n taxa, and suppose that the bipartition b divides the n taxa into two sets with n_0 and n_1 taxa, respectively. The ratio of the number of trees with and without the partition is given by

$$(8) \quad \frac{\left| T(b_i^+) \right|}{\left| T(b_i^-) \right|} = \frac{B(n_0 + 1)B(n_1 + 1)}{B(n) - B(n_0 + 1)B(n_1 + 1)},$$

where, from Felsenstein (1978),

$$(9) \quad B(n) = \frac{(2n-5)!}{(n-3)!2^{n-3}} = \prod_{i=3}^n (2i-5).$$

The formula in equation (8) follows from two facts: 1) the number of possible subtrees one can form on one side of the partition is equivalent to the number of rooted trees for the taxa on that side of the partition, and 2) the number of rooted trees for n taxa is equivalent to the number of unrooted trees for $n+1$ taxa (Felsenstein, 1978).

When we introduce more than one partition factor, the situation becomes more complicated. With k compatible partitions and partition-associated factors, we will be dividing tree space into 2^k classes, each class containing trees with a unique combination of partition factors. For instance, assume that we have five taxa and score the frequency of two compatible bipartitions b_1 and b_2 , which are also the most supported bipartitions. The corresponding factors r_1 and r_2 divide the 15 possible unrooted trees for five taxa into four classes (Table 1). One tree is compatible with both partitions and has the predicted relative probability r_1r_2 . Two trees are compatible with partition 1 but not partition 2, and have predicted relative probability r_1 , whereas two other trees have predicted relative probability r_2 , and are compatible with partition 2 but not with partition 1. Finally, ten trees are not compatible with either partition and have predicted relative probability 1. Because each partition frequency is affected by both factors, we cannot obtain the partition

Table 1. WIB representation of a five-taxon tree space using two multiplicative factors r_1 and r_2 associated with compatible taxon bipartitions.

Number of trees	Compatible with		Predicted relative probability
	partition 1	partition 2	
1	yes	yes	$r_1 r_2$
2	yes	no	r_1
2	no	yes	r_2
10	no	no	1

factors independently from the corresponding partition frequencies. Instead, we need to solve the equation system

$$(10) \quad \frac{s_1}{1-s_1} = \frac{|T(b_1^+, b_2^+)|r_1 r_2 + |T(b_1^-, b_2^-)|r_1}{|T(b_1^-, b_2^+)|r_2 + |T(b_1^-, b_2^-)|}$$

$$(11) \quad \frac{s_2}{1-s_2} = \frac{|T(b_1^+, b_2^+)|r_1 r_2 + |T(b_1^-, b_2^+)|r_2}{|T(b_1^+, b_2^-)|r_1 + |T(b_1^-, b_2^-)|}.$$

Some extra notation will help us generalize this equation system. Suppose that c is a matrix with each row c_i defining a set of fully bifurcating trees by a vector of zeros and ones and recording whether the trees should be consistent with (1) or inconsistent with (0) the corresponding bipartition. That is, if $c_{ij} = 1$, then the trees in the set c_i are consistent with partition b_j ; if $c_{ij} = 0$, then the trees in the set c_i are inconsistent with partition b_j . Furthermore, let $N(c_i)$ be the number of fully bifurcating trees consistent with the specified constraints. Now, obtaining the k partition factors involves solving the equation system

$$(12) \quad \frac{s_1}{1-s_1} = \frac{\sum_{i,c_{i1}=1} (N(c_i) \prod_j r_j^{c_{ij}})}{\sum_{i,c_{i1}=0} (N(c_i) \prod_j r_j^{c_{ij}})}$$

⋮

$$(13) \quad \frac{s_k}{1-s_k} = \frac{\sum_{i,c_{ik}=1} (N(c_i) \prod_j r_j^{c_{ij}})}{\sum_{i,c_{ik}=0} (N(c_i) \prod_j r_j^{c_{ij}})}.$$

The above equation system can be solved numerically using Newton-Raphson's method or the secant method. Table 2 provides some examples of WIB factors calculated for a six-taxon tree to give a feel for the interaction between bipartition factors. Note that the factor associated with a particular support value varies greatly depending on context. If surrounding parts of the tree are well supported, the factor is smaller than if the surrounding parts are poorly supported. Even if all bipartitions have the same support, the factor will vary depending on whether it is associated with a peripheral or a central branch in the tree. The factor will be higher for a central branch simply because the proportion of trees consistent with the partition is smaller. Thus, it takes a stronger data factor to tip the result in favor of trees with such a bipartition. Some results are clearly counterintuitive. For instance, a bipartition frequency of 0.95 is associated with a stronger data effect if the other two partitions have frequencies of 0.70 and 0.95 than if both the other partitions have frequencies of 0.70.

5. A parsimonious excursion

Assume that we find an optimal or near-optimal, fully bifurcating *reference tree*. In the Bayesian context, the reference tree can be the tree with maximum posterior probability (the MAP tree), or it can be obtained by building a fully resolved tree starting with the most frequent bipartition in the MCMC tree sample and then adding compatible bipartitions in order of decreasing frequency, a heuristic procedure that is likely to produce a good estimate of the MAP tree. The WIB factors are calculated from the frequencies of the bipartitions in the reference tree. Now we can understand the WIB method as classifying trees into different groups according to their distance from the reference tree using a single-sided, weighted partition metric (Robinson and Foulds, 1981), in which the weights are the WIB factors of the reference tree rather than the edge lengths of both trees as Robinson and Foulds suggested originally.

Like all partition metrics, WIB representation divides the part of tree space that is close to the reference tree very finely, whereas it loses its resolving power quickly as one moves away from the reference tree. Once all bipartitions in the reference tree have been lost, the WIB representation no longer distinguishes between trees regardless of the fact that they often differ widely both in their resemblance to the reference tree and in their actual posterior probability. This unresolved portion of tree space becomes more and more dominant as the number of taxa increases.

Table 2. Hypothetical examples of partition frequencies (s) and partition factors (r) for an asymmetric six-taxon tree (as in Figure 4a). For each example, the value listed in the middle is the central branch in the tree (i.e., number 2 in Figure 4a). The partition odds (o) are calculated using the formula $o = s / (1 - s)$. The partition factors (r) are given for both the WIB and the WAB methods, and were calculated by solving the appropriate equation system numerically. Finally, upper and lower bounds on the partition factors are given according to formulas discussed in the text. For the upper-bound equation, n is the number of trees without the partition divided by the number of trees with the partition.

s_1-s_3	o	WIB		WAB		Min. $\ln(2o)$	Max. $\ln(no)$
		r	$\ln(r)$	r	$\ln(r)$		
0.70	2.33	6.79	1.92	6.06	1.80	1.54	2.64
0.70	2.33	9.11	2.21	8.32	2.12	1.54	3.21
0.70	2.33	6.79	1.92	6.06	1.80	1.54	2.64
0.80	4.00	10.43	2.34	9.71	2.27	2.08	3.18
0.80	4.00	12.95	2.56	12.20	2.50	2.08	3.75
0.80	4.00	10.43	2.34	9.71	2.27	2.08	3.18
0.90	9.00	20.74	3.03	20.17	3.00	2.89	3.99
0.90	9.00	23.48	3.16	22.90	3.13	2.89	4.56
0.90	9.00	20.74	3.03	20.17	3.00	2.89	3.99
0.95	19.0	40.89	3.71	40.51	3.70	3.64	4.74
0.95	19.0	43.75	3.78	43.37	3.77	3.64	5.31
0.95	19.0	40.89	3.71	40.51	3.70	3.64	4.74
0.80	4.00	11.76	2.46	10.75	2.37	2.08	3.18
0.70	2.33	7.21	1.98	6.95	1.94	1.54	3.21
0.80	4.00	11.76	2.46	10.75	2.37	2.08	3.18
0.90	9.00	26.44	3.27	25.10	3.22	2.89	3.99
0.70	2.33	5.76	1.75	5.72	1.74	1.54	3.21
0.90	9.00	26.44	3.27	25.10	3.22	2.89	3.99
0.95	19.0	55.56	4.02	54.04	3.99	3.64	4.74
0.70	2.33	5.18	1.64	5.17	1.64	1.54	3.21
0.95	19.0	55.56	4.02	54.04	3.99	3.64	4.74
0.80	4.00	12.03	2.49	10.58	2.36	2.08	3.18
0.70	2.33	8.08	2.09	7.62	2.03	1.54	3.21
0.70	2.33	6.65	1.90	6.15	1.82	1.54	2.64
0.90	9.00	27.99	3.33	24.25	3.19	2.89	3.99
0.70	2.33	7.15	1.97	6.96	1.94	1.54	3.21
0.70	2.33	6.50	1.87	6.25	1.83	1.54	2.64
0.95	19.0	60.11	4.10	51.71	3.95	3.64	4.74
0.70	2.33	6.73	1.91	6.65	1.89	1.54	3.21
0.70	2.33	6.42	1.86	6.30	1.84	1.54	2.64

Work on supertree construction using parsimony analysis suggests that there is an alternative factorial representation that extends the resolved part of tree space further away from the reference tree. Much of this work emanates from two papers discussing how phylogenies can be converted into matrices of binary characters for parsimony analysis (Baum, 1992; Ragan,

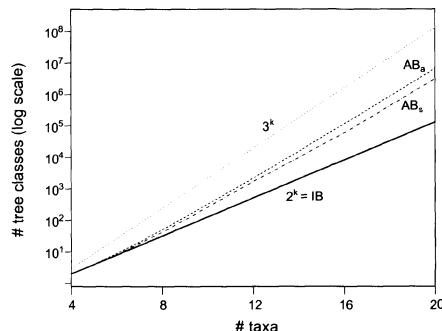


Figure 2. Number of unique tree classes in IB and AB representation. With the IB method, trees are divided into 2^k classes, where k is the number of interior branches in the reference tree. By contrast, the number of AB classes grows almost as fast as 3^k for large trees. The number of unique tree classes is somewhat smaller for symmetric trees (AB_s) than for asymmetric (comb-shaped) trees (AB_a). Because the number of tree classes is given on a log scale, the difference in number of classes between AB and IB representation is dramatic for trees of any decent size.

1992). Both papers arrive at the same method: the reference tree is represented by one binary (two-state) character for each taxon bipartition specified by the tree. This is equivalent to the additive binary coding scheme (Farris *et al.*, 1970) used, among other things, to represent trees in co-evolutionary and biogeographic studies (Brooks, 1981), so let us call it the *Additive Binary* (AB) representation. AB representation is the method commonly referred to as MRP.

The AB matrix has several interesting properties. One of them is that it effectively defines a distance metric on trees based on their parsimony scores. The more distant a tree is to the reference tree, the higher its score will be using the AB matrix. For small tree spaces, the AB representation is equivalent to the unweighted version of the WIB method, which we refer to as *Independent Binary* (IB) representation. The reason is that a binary character can only have the parsimony length of 1 or 2 on trees with five or fewer taxa, and this is equivalent to the presence / absence of the partition under IB representation. For larger tree spaces, however, AB divides the trees more finely than IB using the same number of factors (Figure 2). This is easy to understand in the parsimony framework. Assume that we start with a matrix of binary characters, each representing one of the taxon bipartitions in the reference tree. In other words, each character is equivalent to one of the b vectors discussed previously. Now, the score of a tree under the AB representation is equivalent to the difference in parsimony length between that tree and the reference tree, plus the length of the reference tree. In the IB representation, however, the score is equivalent to the number of

homoplastic characters plus the length of the reference tree. Essentially, the IB factors have an on-off effect, whereas the AB factors show a graded response, the range of which is determined by the way in which the binary character divides the taxa. If the binary character divides the taxon set into subsets with n_0 and n_1 taxa each, then the number of levels of the corresponding AB factor will be $\min(n_0, n_1)$. Not all combinations of AB factor levels are possible, but the AB factors nevertheless divide tree space into many more unique classes than the IB factors (Figure 2). For instance, a 20-taxon tree space will be divided into 130 000 classes by the IB method, but, depending on tree shape, into 3.1 to 6.5 million classes by the AB method.

An obvious improvement of the AB representation is to take the relative support of different taxon bipartitions into account. In such a *Weighted Additive Binary* (WAB) representation, trees that differ from the best tree only in weakly supported parts will be scored as being better than those that do not retain strongly supported groupings. In parsimony analysis, support is commonly measured using either resampling techniques — the bootstrap or the jackknife — or the decay index (Bremer support, branch support). Any of these support measures can be used to calculate the weights of a WAB matrix.

For instance, let us examine the bootstrap-weighting scheme (Ronquist, 1996). Assume that an AB character b_i , which represents the partition i in the best tree, is assigned the weight r_i . Furthermore, assume that the probability of drawing character b_i in bootstrap resampling is r_i / R , where $R = \sum r_i$, and that we count taxon bipartition i as being present in the bootstrap tree when at least one character of type b_i is present in the resampled matrix. Then, the bootstrap support s_i for partition i would be given by the formula $s_i = 1 - (1 - r_i / R)^R$. If R is large, then $(1 - r_i / R)^R \approx e^{-r_i}$ and we can calculate the weight of a particular bipartition from its support value as $r_i = \ln(1 / (1 - s_i))$. If R is small, then r_i can be calculated without using this approximation; for instance, by an iterative method where the initial r_i and R values are obtained using the e^{-r_i} approximation, and then the exact formula is applied repeatedly until the R and r_i values stabilize. When the resultant WAB matrix is subjected to bootstrap resampling, the bootstrap support values of the original analysis should be reproduced accurately.

A more direct relation to the original parsimony scores is obtained in decay-index-derived WAB representation. The decay index d_i for a bipartition i is the difference in parsimony score between the best tree with the partition and the best tree without the partition. The d_i values can be used directly as the weights of the corresponding characters in the WAB matrix. The decay-index-derived WAB scores should be correlated highly with the original parsimony scores, and one might expect the regression coefficient to

be close to 1 because the difference between two trees in WAB scores should reflect the difference between them in parsimony scores. Despite the difference in approach, we might also expect a close correlation between the bootstrap-derived WAB score of a tree and its parsimony length, although the expected value of the regression coefficient is unclear.

To illustrate these different representation methods in the parsimony context, we used a data set of 54 drosophilid sequences of the nuclear protein-coding gene *ADH* (771 bp; Russo *et al.*, 1995). A set of 1000 bootstrap replications was generated in PAUP* v4.0b10 (Swofford, 2002), each analysis using one simple stepwise addition sequence, with *Drosophila persimilis* as the reference taxon, followed by TBR swapping. A fully resolved reference tree was assembled from the bootstrap bipartition frequencies by adding compatible bipartitions in order of decreasing frequency until the tree was fully resolved (only three bipartitions had frequencies less than 0.50). Bootstrap weights of bipartition factors were calculated as $\ln(1 / (1 - (1000s / 1001)))$, as if there had been one extra replication without all bipartitions. Without this correction, it is impossible to calculate the weights of the best supported bipartitions with $s = 1.0$ because $\ln(1 / 1 - s) = \ln(1 / (1 - 1)) = \ln(1/0)$ is undefined. The correction gives a minimum estimate of the strength of the factor associated with these partitions. The decay indices associated with the bipartitions of the reference tree were calculated in a series of constrained searches in PAUP* using the same heuristic strategy as the bootstrap searches.

Finally, we generated 1000 random trees and compared their score using the original data with their score using unweighted, decay-index-weighted, and bootstrap-weighted versions of the IB and AB methods (Figure 3). Because the random trees were distant from the reference tree, the IB and WIB representations failed miserably in predicting tree scores (Figures 3a, c, e). However, all AB and WAB representations provided surprisingly good pictures of tree space, with the weighted versions providing a slight improvement over the unweighted ones (Figures 3b, d, f). It is somewhat surprising that weighting does not improve the AB-based prediction more; however, as one moves closer to the reference tree, the importance of weighting the factors should increase. This is indicated by the results for smaller data sets reported by Ronquist (1996), in which random trees were closer to the reference tree (Table 3). The IB and WIB methods were not included in this study, but one would expect the WIB representation to outperform IB as well as AB representation (equivalent to MRP) in the region close to the reference tree because the latter representation will be similar to IB in this part of tree space.

Thus, given the success of the WAB method in the parsimony context, is it possible to use it for describing Bayesian tree probabilities?

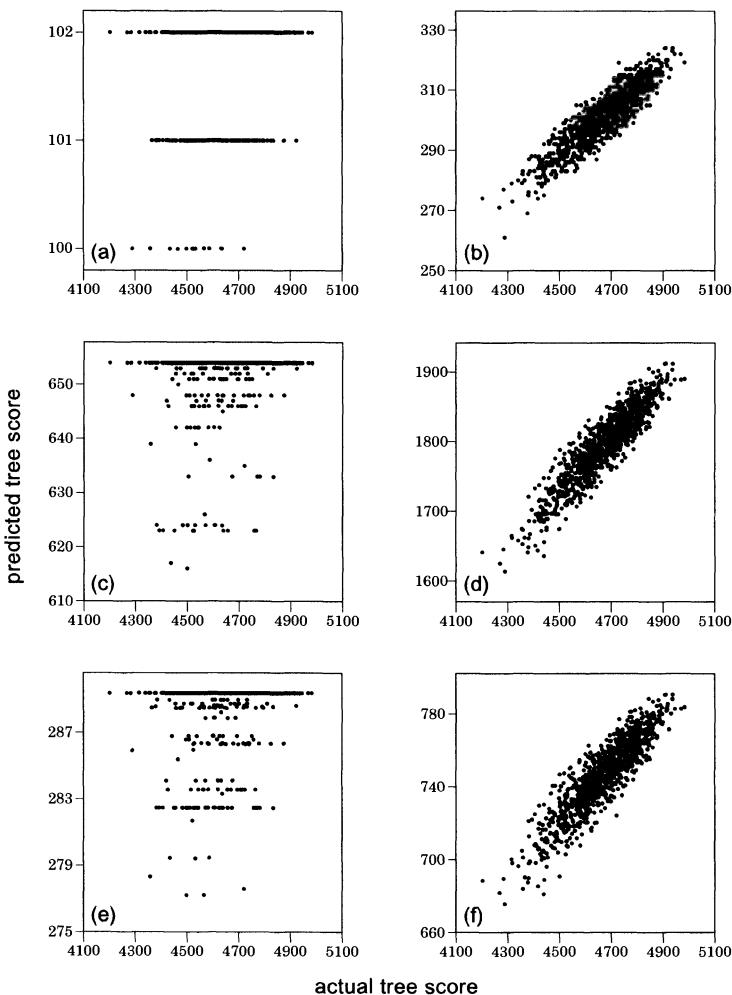


Figure 3. Performance of different representation methods in predicting the parsimony scores of 1000 random trees for a 54-taxon data set of drosophilid *ADH* sequences: a) IB representation, b) AB representation, c) decay-index-weighted WIB representation, d) decay-index-weighted WAB representation, e) bootstrap-weighted WIB representation, and f) bootstrap-weighted WAB representation.

Table 3. Performance of AB, decay-index-weighted WAB (WAB_d), and bootstrap-weighted WAB (WAB_b) representation in predicting parsimony tree scores of 1000 random trees for some different morphological data sets (modified from Ronquist, 1996).

Data set	Taxa	Characters	Coefficient of determination (r^2)		
			AB	WAB_d	WAB_b
Liopteridae	11	54	0.833	0.933	0.929
Cynipidae 2	12	108	0.889	—	0.852
Ibaliidae	18	82	0.686	0.922	0.912
Cynipoidea	19	110	0.867	0.933	0.929
Cynipidae 2	32	158	0.642	—	—

6. Weighted additive binary representation

Implementing WAB representation in the Bayesian framework is actually straightforward: we just translate each parsimony step in a WAB factor to a constant difference in log-likelihood scores between trees. For instance, assume that the binary parsimony character associated with a WAB factor r can have one, two, or three steps on different trees. The trees with two steps are then r times as likely as those with three steps. Similarly, the trees with one step are r times more likely than those with two steps and r^2 times more likely than those with three steps.

To solve for WAB instead of WIB factors, we make the following changes to our previous notation. Assume that $l(b_i, c_i)$ is the parsimony length of the binary character b_i on trees in the set defined by c_i and that $g(b_i)$ is the maximum parsimony length of character b_i on any tree. Now, let c be the matrix containing all unique rows c_i , where each element $c_{ij} \in (0, \dots, g(b_j))$. Each row c_i will then define a set of trees where $l(b_j, c_i) = g(b_i) - c_{ij}$. In some cases, the set will be empty; these rows can be deleted from c for more efficient calculation. With these changes, we can solve for the WAB-partition factors using the same equation system defined above for the WIB factors (equations (12) to (13)). The only difference is that we must account for the fact that the elements of c might be greater than 1. The equation system thus becomes

$$(14) \quad \frac{s_1}{1-s_1} = \frac{\sum_{i, c_{i1} \geq 1} \left(N(c_i) \prod_j r_j^{c_{ij}} \right)}{\sum_{i, c_{i1} = 0} \left(N(c_i) \prod_j r_j^{c_{ij}} \right)}$$

⋮

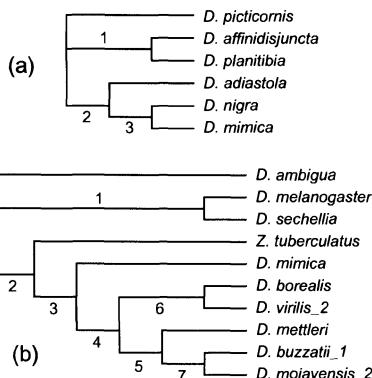


Figure 4. Subsets of the drosophilid data set used to examine factorial representation of Bayesian tree probabilities: (a) a six-taxon subset representing a small subsection of the best tree for the full data set, and (b) a ten-taxon subset spanning the entire best tree for the full data set. The taxon bipartition numbers are the same as those in Table 4.

$$(15) \quad \frac{s_k}{1-s_k} = \frac{\sum_{i, c_{ik} \geq 1} \left(N(c_i) \prod_j r_j^{c_{ij}} \right)}{\sum_{i, c_{ik} = 0} \left(N(c_i) \prod_j r_j^{c_{ij}} \right)}.$$

Note that there will be more rows in the c matrix when we solve for WAB factors, reflecting the larger number of tree classes, but the number of WAB factors is still k .

Asymmetric six-taxon trees are the smallest reference trees for which WIB- and WAB-representation methods differ: WIB representation results in eight tree classes, whereas WAB representation results in nine. Examination of hypothetical support values for such trees (Table 2) reveals that WIB and WAB factors are generally similar, although not identical. When they differ, WAB factors are smaller than WIB factors. As the support level increases, the WIB and WAB factors become more similar.

To illustrate the efficiency of WIB and WAB representation in describing empirical Bayesian tree probabilities, we selected a six-taxon and a ten-taxon subset of the 54-taxon drosophilid *ADH* data set (Figure 4). For each data set, the best (MAP) tree was estimated through MCMC sampling (1 000 000 generations, sampled every 100th generation, first 50% of samples discarded as burn-in) under the GTR model with site-specific rates using MrBayes 3.0b4 (Ronquist and Huelsenbeck, 2003). Default values in MrBayes 3.0b4 were used for priors and other settings. For both data sets, all taxon bipartitions in the MAP tree were supported strongly.

For the six-taxon data set, there are 105 possible unrooted trees, which the WIB and WAB methods divide into eight and nine classes, respectively, based on the reference tree. For the ten-taxon data set, there are about 2×10^6 different trees, which are divided into 128 (WIB) or 290 (WAB) classes based on the reference tree. For the six-taxon tree, we estimated the marginal posterior probability of each of the 105 possible trees using the harmonic mean of the likelihood values (Newton *et al.*, 1994; for a discussion of alternative estimators of marginal likelihood values, see Gamerman, 1997) by feeding the program with the appropriate starting tree and setting the proposal probability of all topology proposals to zero. The estimated tree probabilities were used, in turn, to estimate the odds $o_i = s_i / (1 - s_i)$ for each taxon bipartition. For the ten-taxon problem, we used a similar approach, except that we selected one tree randomly from each of the 290 WAB classes and used only 500 000 generations in each topology-constrained MCMC analysis. To increase the accuracy of the estimates of bipartition odds, we also estimated the marginal posterior probabilities for all trees differing from the reference tree with a WAB score of two or less, and included them in the estimate together with the 290 random trees, deleting duplicates. The bipartition odds were used to obtain the corresponding WIB and WAB factors by solving the appropriate equation system (i.e., (12) to (13)) using the secant method (Table 4). Finally, the estimated posterior tree probabilities were plotted against the tree probabilities predicted by the WIB and WAB methods (Figure 5).

For the six-taxon tree space, the WIB and WAB methods were very similar (Figures 5a, b). The WAB method split the worst trees into two classes, making the predicted probabilities match the actual probabilities better. The best trees had a probability that matched the predicted value closely (line in diagram). For the poor trees, the maximum probability of trees in each class was predicted more accurately than the average probability.

For the ten-taxon trees, we saw much more of a difference between WIB and WAB representations (Figures 5c, d). Both methods predicted the near-optimal trees accurately and overestimated the probability of poor trees. However, the WIB prediction degraded quickly as we moved away from the reference tree, whereas the WAB prediction remained distinctly correlated with the actual tree probability all the way to the trees that were most distant from the reference tree. Given that there are about two million trees in this tree space, it is remarkable that the WAB method managed to capture so much of its properties with only seven weighted factors.

Is the increased power of the WAB method important in the Bayesian context? There are several reasons to suggest that the answer is yes. First, if there is significant conflict between the data set at hand and a prior based on

Table 4. Measures of support for the six-and ten-taxon examples (Figures 4 and 5). The partition odds (o) were estimated from marginal tree probabilities as described in the text, and the partition factors (r) were calculated from these by solving the appropriate equation system. For these well-supported example trees, the WIB and WAB factors (r) are identical to the precision reported here. As predicted, the factors are higher than the corresponding partition odds. Two methods of estimating the factors are given. One is based on taking twice the partition odds ($2o$), which provides a lower bound on the factors. The other method is based on the harmonic mean (h) of the estimates from the two near-optimal trees breaking only that partition. Both methods provide accurate approximations of the partition factors.

Partition	$\ln(o)$	$\ln(r)$	$\ln(2o)$	$\ln(h)$
Six-taxon tree				
1	4.76	5.45	5.45	5.45
2	6.89	7.62	7.58	7.56
3	4.08	4.78	4.78	4.77
Ten-taxon tree				
1	32.77	33.46	33.46	33.75
2	73.89	74.58	74.58	75.81
3	31.82	32.52	32.51	32.51
4	5.85	6.54	6.54	6.65
5	9.25	9.96	9.94	9.94
6	45.60	46.30	46.29	46.32
7	4.65	5.34	5.34	5.38

the results from a previous analysis, then it becomes important to know what the previous analysis implied concerning the relative probabilities of trees that were poor in that analysis. Second, as trees become larger, the portion of tree space resolved by the WIB method becomes smaller. For large trees, the portion of tree space resolved by the WIB method might be simply too small for an appropriate description of the probability distribution on trees. Third, an MCMC sampler needs to move across topology space to find all the trees it should sample from. The presence / absence nature of the WIB factors might cause artificial threshold effects that prevent an MCMC sampler from moving around in tree space as it should.

7. Estimating bipartition factors

Solving the equation system (12) to (13), even with iterative numerical methods, is feasible only for relatively small trees. Thus, it is desirable to develop more computationally efficient ways of estimating the WIB and WAB partition factors.

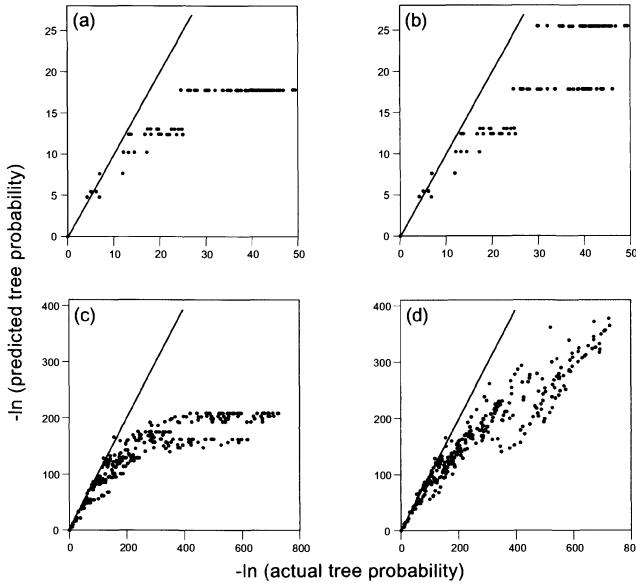


Figure 5. Factorial representation of Bayesian tree spaces for a six- and a ten-taxon (Figures 4a, b, respectively) example: a) WIB representation of the six-taxon tree space, b) WAB representation of the six-taxon tree space, c) WIB representation of the ten-taxon tree space, and d) WAB representation of the ten-taxon tree space. The line marks perfect prediction of tree probabilities. All tree probabilities are measured relative to the best tree.

7.1 Upper and lower bounds

One can find upper and lower bounds on the WIB factors quickly and easily by focusing on the bipartition-odds equation,

$$(16) \quad n = \frac{\left| T(b^-_1) \right|}{\left| T(b^+_1) \right|} \frac{s_1}{1 - s_1},$$

which is valid when there is only one factor affecting tree probabilities. The equation says that the partition factor r_1 equals the ratio of inconsistent to consistent trees times the partition odds (ratio of samples with and without the partition). The fewer the consistent trees are in number relative to the inconsistent trees, the stronger the partition factor must be to achieve the observed partition frequency. As we have seen, the ratio of inconsistent to

consistent trees is determined by the number of taxa in the two partitions according to equations (8) and (9).

If we add a second taxon bipartition that is compatible with the first partition, with the associated factor $r_2 > 1$, we can use a similar equation for r_1 , but we need to take the altered probability of the other trees into account. Then we have

$$(17) \quad r = \frac{|T(b_1^-, b_2^+)|r_2 + |T(b_1^-, b_2^-)|}{|T(b_1^+, b_2^+)|r_2 + |T(b_1^+, b_2^-)|} \frac{s_1}{1-s_1}.$$

How the addition of r_2 affects the value of r_1 depends on the relation between the ratio of the numbers we multiply with r_2 ($|T(b_1^-, b_2^+)| / |T(b_1^+, b_2^+)|$) and the original ratio ($|T(b_1^-)| / |T(b_1^+)|$) (see equation (16)). If the ratio of numbers multiplied by r_2 is smaller than the original ratio, the value of r_1 will decrease. Because

$$(18) \quad \frac{|T(b_1^-, b_2^+)|}{|T(b_1^+, b_2^+)|} < \frac{|T(b_1^-)|}{|T(b_1^+)|},$$

which can be shown using equations (8) and (9) to specify the number of trees satisfying each of the constraints, then r_1 will always decrease when r_2 is added. A similar argument can be used to show that r_1 decreases when the next congruent partition is added, and so on. Therefore, the estimate of r_1 decreases monotonously as more and more compatible partitions are added to the equation, and, hence, equation (16) provides an upper bound on a WIB factor; that is,

$$(19) \quad r \leq \frac{|T(b_i^-)|}{|T(b_i^+)|} \frac{s_i}{1-s_i}.$$

A lower bound on WIB factors is obtained by considering the case when all other possible taxon bipartitions are supported so strongly that we can ignore the trees inconsistent with them. Then we need to consider only the three remaining alternative trees centered on the taxon bipartition in focus, one of which is consistent with the bipartition and two of which are not. In other words, the lower bound on a WIB factor is

$$(20) \quad r_i \geq \frac{2s_i}{1 - s_1}.$$

Under similar conditions (all partitions are compatible), the lower and upper bounds of WIB factors apply to WAB factors as well. Actually, unlike WIB factors, it appears possible to derive a lower upper bound on WAB factors, but this is not attempted here. In practice, the real WIB or WAB factors are likely to be much closer to the lower than the upper bound, particularly in well-supported trees. This occurs because the upper bound is associated with trees in which all nodes except the one being examined have frequencies close to zero (given that the tree is not extremely small). Even a weakly supported tree is unlikely to be close to this situation. Thus, the lower-bound equation can be used as a good-and-fast approximation of the real partition factors. For example, for the well-supported six- and ten-taxon trees examined above, the lower-bound equation provides a very accurate approximation of the partition factors (Table 4). In trees with less strongly supported partitions, the partition factors will be considerably higher than the lower bound, but still far removed from the upper bound (Table 2).

7.2 Estimating partition odds directly

A problem with the lower-bound approximation is that the partition odds on which it is based are difficult to estimate for well-supported partitions. When a partition frequency approaches 1.0, the ordinary estimate of the odds (i.e., the ratio of samples with and without the partition) becomes unstable because there will be few MCMC samples without the partition. Indeed, one encounters frequently partitions that are present in all MCMC samples, in which case this estimate becomes undefined (division by zero). A possible approach then is to estimate the odds using $n / 1$, where n is the number of MCMC samples, as if one had observed one sample without the partition. This solution is analogous to the one used above in calculating bootstrap-derived weights for WAB or WIB characters. An obvious disadvantage of this method is that it puts an artificial cap on the values of partition factors, which will distort the factorial representation of tree probabilities unnecessarily. Furthermore, the method does not address the difficulty of estimating partition odds accurately when there are only few sampled trees conflicting with the partition. For instance, one or two sampled trees without the partition translates to a difference in estimated odds by a factor of two, which becomes important when one realizes that stochastic effects might well result in some runs giving one tree without the partition, whereas others will give two.

A better (although computationally more complex) approach to estimating extreme partition odds is to evaluate the marginal likelihood of the alternative hypotheses using separate MCMC analyses, one constrained to sample from trees with the partition and the other one constrained to sample from trees without it. There are several estimators of the marginal likelihood that can be used for this purpose, of which the harmonic mean of the likelihood values used above (Newton *et al.*, 1994) is the easiest one to implement. In our experience, this estimator is stable enough to provide a rough approximation of the likelihood ratio of the contrasting hypotheses. A more accurate estimate can be obtained in a subsequent analysis by increasing the likelihood of the trees inconsistent with the examined bipartition artificially such that the two alternative hypotheses can be contrasted in a single MCMC run.

7.3 Estimating partition factors directly

Similar methods can be used to estimate partition factors directly without calculating the partition odds first. Any pair of trees that differ from each other solely by a single step in one WAB factor could be used in estimating the magnitude of that factor because the factor is a predictor of the likelihood ratio of the trees. One way of utilizing this fact would be to estimate the partition factors in the reference tree by a series of searches, in each of which all bipartitions of the reference tree are fixed except one. This would leave only three alternative topologies (one consistent and two inconsistent with the bipartition) to be considered in estimating each partition factor. If the partition factor is weak, the relative probability of the three competing trees can be evaluated in a single run; if it is strong, one might have to estimate the likelihood scores of the two topologies that conflict with the partition in separate MCMC analyses or inflate their probability artificially to make accurate estimation of the factor possible in a single analysis, as discussed above for the estimation of partition odds. Because the topology space is so small for these analyses, they should be much less demanding computationally than those required in estimating partition odds. Our empirical results suggest that the harmonic mean of the estimates based on the two inconsistent trees comes very close to the true partition factor (Table 4).

The method could be refined by examining two or more adjacent partitions at a time. The estimate of each partition factor would then be based on comparisons among more trees than the three used in the original formulation, but the computational burden would, of course, increase correspondingly.

8. Partition factors: a better support measure?

The method of estimating partition factors suggests a similarity between partition factors and the probabilistic analog of the decay index. In the Bayesian context, the decay index would be the ratio between the marginal probabilities of the best trees with and without the partition. In many cases, the first tree would be the same as the reference tree and the second would differ only in that it was incongruent with the partition in question. Thus, the Bayesian decay index would be the same in many cases as selecting the minimum, rather than calculating the harmonic mean, of the two estimates of the partition factor based on near-optimal trees.

To the extent that they differ, however, the partition factors should describe the partition-associated data effects more accurately than the decay indices, which will underestimate the effects systematically. In a sense, the partition factors provide the ultimate support measure because they are related to partition-associated data effects more intimately than both the decay index and the bipartition frequencies. The correlation between bipartition frequencies, in particular, and data effects is imprecise. A frequency of, say, 0.7 indicates considerable data support if it is associated with a partition in the middle of a poorly supported tree, but it implies only modest data support when tied to a peripheral partition in a well-supported tree. These effects are illustrated by the previously discussed hypothetical examples for a six-taxon data set (Table 2). In these, a taxon-bipartition frequency of 0.7 can be associated with a partition factor ranging from 5.2 to 8.3 depending on the context. Similarly, a taxon-bipartition frequency of 0.95 can translate to a partition factor ranging from 40.5 to 54.0. Larger trees will display even more variability in the partition factors associated with a given partition frequency.

Because the partition factors hold the key to the distribution of tree probabilities, which plays such a central role in Bayesian phylogenetic inference and supertree construction, it would be desirable if future Bayesian analyses reported partition factors in addition to, or even instead of, partition frequencies.

9. Using partition factors to build supertrees

Using the WAB partition factors from one analysis to define a prior-probability distribution on trees for a subsequent Bayesian MCMC analysis is straightforward. As the MCMC procedure of the second analysis moves around among possible trees, it needs to evaluate the prior-probability ratio between trees from the partition factors. This is done easily by evaluating the

difference in the WAB parsimony score of the trees using parsimony algorithms. The binary data matrix used in this step is exactly the same as the matrix used in MRP, except that the characters are weighted differently. Once the prior ratio is obtained, the likelihood ratio of the trees is calculated using standard procedures and the MCMC analysis proceeds as usual. A possible complication that might occur is that the WAB representation introduces some artificial step effects in the distribution of prior tree probabilities, which might cause difficulties for the MCMC procedure. Only empirical studies can determine whether this is an issue in real analyses, and this is outside of the scope of the present chapter. However, we note that the improved resolution of WAB representation over WIB representation should help to alleviate this problem. Similar considerations suggest that Bayesian analyses constrained by a single partition factor should use WAB rather than WIB in translating the partition factor to prior tree probabilities.

Building a composite prior from several previous analyses only involves combining several WAB matrices. If there is partial rather than complete taxon overlap among the previous analyses, the WAB characters are simply coded as missing observations for the taxa that are not represented in the reference tree of a particular analysis. This results in the position of the missing taxa having no effect on the tree probabilities from that analysis, as is appropriate. The previous analyses need not be Bayesian as long as the results can be interpreted in terms of probabilities. For instance, the bootstrap proportions from a parsimony analysis could be treated simply as if they were Bayesian clade probabilities. This should result in a reasonable prior, given that the parsimony criterion is consistent for the type of data analyzed. Even a tree presented without support values could be accepted as providing some prior evidence concerning topology. One possible way of converting such a tree into probabilities would be to assume that each clade (taxon bipartition) was supported by one character in a parsimony analysis, corresponding to a bootstrap proportion of roughly 0.63 ($\ln(1 / (1 - 0.63)) \approx 1.0$). The independence of previous analyses is an additional problem that should be considered (see also Gatesy and Springer, 2004). Clearly, it is inappropriate to multiply the probabilities together from two previous analyses if they are based on the same or very similar data. Expressing prior knowledge in terms of probabilities is difficult and it is important to describe the approach taken clearly and to examine the robustness of the conclusions to alternative ways of formulating the prior. As synthesis becomes increasingly important in phylogenetics, however, it is a challenge that has to be faced.

Constructing a Bayesian supertree is the same thing as sampling from a composite prior assembled by combining WAB matrices without adding any new data. As long as the WAB matrices provide efficient descriptions of the

Bayesian tree probabilities and the underlying analyses and models are independent (see below), the supertree space is fully specified statistically. Perhaps the most intriguing aspect is that the supertree space can be sampled by an MCMC procedure considerably faster and more intelligently than an ordinary tree space. Sampling from an ordinary tree space involves calculating the likelihood ratios of trees, a computationally expensive operation that is responsible for the bulk of the execution time of a Bayesian MCMC analysis. Sampling from a supertree space requires only evaluation of the difference between two trees in weighted parsimony scores, a vastly more efficient operation. Using published shortcuts (Goloboff, 1996; Ronquist, 1998), this can be completed in time dependent linearly on the number of WAB factors, but independent of supertree size. Even more intriguing, the supertree space can be sampled using fully conditional (Gibbs) sampling. In normal MCMC sampling of topologies (using the Metropolis algorithm), new trees are proposed according to a proposal distribution and then evaluated according to their posterior probability. Because of the difficulties of formulating a good proposal distribution, most proposed trees have low posterior probability and are rejected. This leads to computational inefficiency; it is not unusual to see 90% or more of topology proposals being rejected, making it necessary to run the MCMC procedure much longer than otherwise necessary. In Gibbs sampling, however, the proposal distribution is the same as the posterior-probability distribution conditional on the present state (Gelman *et al.*, 1995). This results in automatic acceptance of all proposals and more efficient MCMC sampling. In the Bayesian supertree context, one could implement Gibbs sampling for topologies easily. For instance, one could propose a new tree by erasing a few adjacent nodes in the existing tree, then calculate the parsimony score (log probabilities) of all possible ways of reconstructing a fully bifurcating tree, and finally select among these according to their relative probabilities.

Given these considerations, it is clear that the construction of a Bayesian supertree is comparable in computational complexity to the construction of a supertree using parsimony analysis of weighted or unweighted MRP matrices. Indeed, there are reasons to expect the Bayesian MCMC approach to be more efficient than parsimony. First, sampling a tree space adequately can be easier than exploring large islands of optimal or near-optimal trees. Second, the random component of MCMC sampling often appears to provide an efficient way of climbing from trees with poor likelihood scores to trees with optimal or near-optimal scores without becoming trapped in local optima. These advantages come in addition to the fact that the Bayesian supertree space is based on fully specified evolutionary models, with all the benefits associated with such an approach.

10. Model independence assumption

Now it is time to return finally to the assumption made at the beginning of this chapter, namely that we could focus entirely on the topology parameter and ignore the other parameters in the phylogenetic model. As stated then, the phylogenetic model typically includes branch length and substitution model parameters in addition to topology. When a diffuse or non-informative prior is formulated, each parameter is associated typically with a completely independent probability distribution. Thus, the joint probability of a particular topology and a particular set of branch lengths, for example, is calculated by simply multiplying the probability of the topology and the probability of the branch lengths. When the posterior is obtained, it is not solely a probability distribution on trees; it is a joint-probability distribution on all the parameters in the model. The posterior tree probabilities we have been discussing so far are marginal probabilities that are obtained by integrating and summing out other parameters in the phylogenetic model from the joint posterior-probability distribution.

What happens when we embark on a subsequent analysis and want to use the posterior distribution from the first analysis as the prior in the next? If the non-topology parameters for the new data set are entirely independent from those of the old data set, we can use the marginal tree probabilities without any problems. This involves assuming that, for instance, the branch-length parameters are independent among partitions. If, some of the non-topology parameters are shared, however, we should take this into account. Unfortunately, this is rather complicated because of the dependence among parameters. In particular, branch-length parameters are likely to be highly dependent on topology. For instance, the joint posterior-probability distribution from the first analysis might specify that the probability of a certain branch (corresponding to a certain taxon bipartition) being of length v is p_1 on topology 1 and p_2 on topology 2. Similarly, if the branch is of length v_1 , topology 1 might be p_1 more probable than topology 2, whereas a length of v_2 implies that topology 1 is p_2 more probable than topology 2.

There are at least two possible ways to address this dependence among parameters. One is to ignore that the non-topology parameters are shared across data sets, which will obviously result in some loss of information. If the only effect is to make the probability distribution on trees slightly more diffuse, however, this might be a cheap price to pay for the ability to build supertrees. Another simplifying approximation, which will work for some parameters but not for branch lengths, is to ignore the dependence and use the marginal posterior distributions as truly independent priors. For instance, the posterior-probability distributions of GTR rate parameters or stationary-

state frequencies might often be sufficiently robust to changes in topology that the dependence between these distributions can be ignored safely.

In conclusion, then, it might often be justifiable to use the marginal posterior tree probabilities of one Bayesian analysis as an independent prior-probability distribution on trees in a subsequent analysis, ignoring model parameters shared across data sets.

11. Synthesis

Supertree construction is viewed usually as a procedure that takes trees resulting from individual phylogenetic analyses and puts them together into a larger tree. The Bayesian approach, with its emphasis on probability distributions, does not fit easily into this scheme. The result of Bayesian phylogenetic inference is not a single tree, but a distribution of tree probabilities. To build a true Bayesian supertree, we need to combine probability distributions on trees, not single trees.

As we have seen, tree probability distributions are difficult to handle because of the enormous size of tree space. To work with tree probability distributions, we need efficient ways of describing them and taxon-bipartition frequencies have become the standard tool in Bayesian phylogenetic inference. The partition frequencies are statements about the relative probability of different classes of trees, but they do not allow direct calculation of individual tree probabilities. To accomplish that, we have described two ways of translating partition frequencies to approximate tree probabilities. Both rely on the assumption that there are independent partition-associated multiplicative data effects. In one method (WIB), each data effect is either present or absent depending on whether or not the taxon bipartition is represented in the tree. In the other (WAB), there is a graded response depending on the extent to which the tree conflicts with the taxon bipartition represented by the data effect. The level of the effect is determined by the parsimony score of a binary (MRP) character describing the partition. We have seen that the latter approach is superior in describing the tree-probability distribution, particularly for that part of tree space far removed from the best trees. We have also seen that the WAB data factors can be calculated rapidly from partition frequencies using the lower-bound estimate or estimated more accurately using specifically designed Bayesian MCMC analyses. WAB matrices from individual Bayesian analyses can be used to specify a prior-probability distribution on subsequent analyses. By combining several WAB matrices, we can define a supertree space that has special properties that reduce the computational complexity of sampling from it with Bayesian MCMC methods.

We think that the field of supertree construction in general would benefit from more of a Bayesian perspective on phylogenetic inference. Instead of viewing phylogenetic methods as producing a single tree, we should focus on the way they translate data sets to distributions of tree scores (parsimony scores, maximum likelihood scores, bootstrap proportions). Just picking a single tree from each analysis amounts to discarding most of the information in the data. MRP is often regarded as a way of coding a single tree for inclusion in a supertree, but it is also an excellent way of summarizing tree scores. The tree used to derive the MRP characters is just a vehicle for finding the significant partition-associated data effects. As shown here and elsewhere, MRP is a surprisingly efficient key to the entire distribution of tree scores in the original analysis, particularly if the MRP characters are weighted differentially (Ronquist, 1996; Bininda-Emonds and Sanderson, 2001). The success of bootstrap-derived and, in particular, decay-index-derived weights in improving MRP prediction of tree scores can be understood easily because of their similarity to WAB partition factors. Indeed, the Bayesian WAB perspective suggests new and efficient ways of calculating MRP weights.

There has been much emphasis lately on the advantages of polynomial-time algorithms for constructing supertrees (e.g., Semple and Steel, 2000; Steel *et al.*, 2000; Page, 2002, 2004). However, the polynomial-time algorithms (e.g., MINCUTSUPERTREE; Semple and Steel, 2000) that have been discussed so far in the supertree context are consensus-like techniques that typically take one tree from each analysis as input and, therefore, ignore the distribution of tree scores. They can be improved by inputting sets of weighted trees from each analysis, but even a large set of weighted trees is a far less efficient way of representing the distribution of tree scores than factorial methods such as WAB. Ultimately, the choice of supertree method must depend on a balance between speed and accuracy. Factorial methods have the upper hand when it comes to accuracy, and supertrees can be constructed undoubtedly from factorial supermatrices using standard polynomial-time tree construction algorithms. We suspect, however, that many systematists will prefer more powerful supertree inference techniques.

What is the computational complexity of building fully Bayesian supertrees based on MCMC sampling from tree spaces defined by WAB supermatrices? As we have tried to show, there is nothing to suggest that these analyses would be more complex computationally than parsimony analyses. Given that the parsimony method has been applied successfully to problems with thousands of sequences, it should be possible to construct Bayesian supertrees of at least this size. By being based on fully specified evolutionary models, the Bayesian method potentially could make more

efficient use of the underlying information than all alternative supertree methods available today.

What developments are necessary to enable the construction of Bayesian supertrees on a grand scale? Not much. Reporting taxon-bipartition frequencies is common practice already in Bayesian phylogenetic inference, and the frequencies can be converted rapidly to partition factors using the approximation based on the lower-bound equation as described above. However, as discussed in detail in this chapter, it is possible to estimate the partition factors more accurately, and it would be advantageous if it became standard practice to report such partition-factor estimates in addition to the partition frequencies in Bayesian analyses. Constructing an MCMC algorithm that samples from a supertree space defined by a WAB supermatrix is straightforward. In the near future, we hopefully will see the first Bayesian supertrees reported and learn more about the potential of this approach to supertrees.

Acknowledgements

We would like to thank Olaf Bininda-Emonds for inviting us to write this chapter and for being patient with its delivery. Lars Cederberg performed some of the analyses. Sverker Holmgren, Karljohan Lundin, and Lars Gustavsson provided invaluable help in developing numerical methods for calculating partition factors from taxon bipartition frequencies. Olaf Bininda-Emonds, Rutger Vos, and an anonymous reviewer provided comments that helped improve the original version of the chapter. FR was supported by the Swedish Research Council (grant 621–2001–2963) and JPH by the National Science Foundation (grants DEB–0075406 and MCB–0075404).

References

- BAUM, B. R. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon* 41:1–10.
- BAUM, B. R. AND RAGAN, M. A. 2004. The MRP method. In O. R. P. Bininda-Emonds (ed). *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 17–34. Kluwer Academic, Dordrecht, the Netherlands.
- BININDA-EMONDS, O. R. P. AND BRYANT, H. N. 1998. Properties of matrix representation with parsimony analyses. *Systematic Biology* 47:497–508.
- BININDA-EMONDS, O. R. P. AND SANDERSON, M. J. 2001. Assessment of the accuracy of matrix representation with parsimony analysis supertree construction. *Systematic Biology* 50:565–579.

- BROOKS, D. R. 1981. Hennig's parasitological method: a proposed solution. *Systematic Zoology* 30:229–249.
- FARRIS, J. S., KLUGE, A. G., AND ECKHARDT, M. J. 1970. A numerical approach to phylogenetic systematics. *Systematic Zoology* 19:172–191.
- FELSENSTEIN, J. 1978. The number of evolutionary trees. *Systematic Zoology* 27:27–33.
- GAMERMAN, D. 1997. *Markov Chain Monte Carlo*. Chapman and Hall, Boca Raton, Florida.
- GATESY, J. AND SPRINGER, M. S. 2004. A critique of matrix representation with parsimony supertrees. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 369–388. Kluwer Academic, Dordrecht, the Netherlands.
- GELMAN, A., CARLIN, J. B., STERN, H. S., AND RUBIN, D. B. 1995. *Bayesian Data Analysis*. Chapman and Hall, Boca Raton, Florida.
- GOLOBOFF, P. A. 1996. Methods for faster parsimony analysis. *Cladistics* 12:199–220.
- HUELSENBECK, J. P., RONQUIST, F., NIELSEN, R., AND BOLLBACK, J. P. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 294:2310–2314.
- HUELSENBECK, J. P., LARGET, B., MILLER, R. E., AND RONQUIST, F. 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. *Systematic Biology* 51:673–688.
- LEWIS, P. O. 2000. Phylogenetic systematics turns over a new leaf. *Trends in Ecology and Evolution* 16:30–37.
- NEWTON, M. A., RAFTERY, A. E., DAVISON, A. C., BACHA, M., CELEUX, G., CARLIN, B. P., CLIFFORD, P., LU, C., SHERMAN, M., TANNER, M. A., GELFAND, A. E., MALICK, B. K., GELMAN, A., GRIEVE, A. P., KUNSCH, H. R., LEONARD, T., HSU, J. S. J., LIU, J. S., RUBIN, D. B., LO, A. Y., LOUIS, T. A., NEAL, R. M., OWEN, A. B., TU, D. S., GILKS, W. R., ROBERTS, G., SWEETING, T., BATES, D., RITTER, G., WORTON, B. J., BARNARD, G. A., GIBBENS, R., AND SILVERMAN, B. 1994. Approximate Bayesian inference by the weighted bootstrap (with discussion). *Journal of the Royal Statistical Society, Series B* 56:3–48.
- PAGE, R. D. M. 2002. Modified mincut supertrees. In R. Guigó and D. Gusfield (eds), *Algorithms in Bioinformatics, Second International Workshop, WABI 2002, Rome, Italy, September 17–21, 2002, Proceedings*, pp. 537–552. Springer, Berlin.
- PAGE, R. D. M. 2004. Taxonomy, supertrees, and the Tree of Life. In O. R. P. Bininda-Emonds (ed.). *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 247–265. Kluwer Academic, Dordrecht, the Netherlands.
- PURVIS, A. 1995. A modification to Baum and Ragan's method for combining phylogenetic trees. *Systematic Biology* 44:251–255.
- RAGAN, M. A. 1992. Phylogenetic inference based on matrix representation of trees. *Molecular Phylogenetics and Evolution* 1:53–58.
- ROBINSON, D. F. AND FOULDS, L. R. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences* 53:131–148.
- RONQUIST, F. 1996. Matrix representation of trees, redundancy, and weighting. *Systematic Biology* 45:247–253.
- RONQUIST, F. 1998. Fast Fitch-parsimony algorithms for large data sets. *Cladistics* 14:387–400.
- RONQUIST, F. AND HUELSENBECK, J. P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572–1574.
- RUSSO, C. A. M., TAKEZAKI, N., AND NEI, M. 1995. Molecular phylogeny and divergence times of drosophilid species. *Molecular Biology and Evolution* 12:391–404.
- SANDERSON, M. J., PURVIS, A., AND HENZE, C. 1998. Phylogenetic supertrees: assembling the trees of life. *Trends in Ecology and Evolution* 13:105–109.

- SEMPLE, C. AND STEEL, M. 2000. A supertree method for rooted trees. *Discrete Applied Mathematics* 105:147–158
- STEEL, M., DRESS, A., AND BÖCKER, S. 2000. Simple but fundamental limitations on supertree and consensus tree methods. *Systematic Biology* 49:363–368.
- Swofford, D. L. 2002. *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Version 4. Sinauer, Sunderland, Massachusetts.

3. Methodological considerations

Chapter 10

SOME DESIDERATA FOR LIBERAL SUPERTREES

Mark Wilkinson, Joseph L. Thorley, Davide Pisani, François-Joseph Lapointe, and James O. McInerney

Abstract: Although a variety of supertree methods have been proposed, our understanding of these methods is limited. In turn, this limits the potential for biologists who seek to construct supertrees to make informed choices among the available methods. In this chapter, we distinguish between supertree methods that offer a conservative synthesis of the relationships that are agreed upon or uncontradicted by all the input trees and liberal-supertree methods that have the potential to resolve conflict. We list a series of potential desirable properties (“desiderata”) of liberal-supertree methods, discuss their relevance to biologists, and highlight where it is known that particular methods do or do not satisfy them. For biologists, the primary aim of liberal-supertree construction is to produce accurate phylogenies and most of our desiderata relate to this prime objective. Secondary desiderata pertain to the practicality of supertree methods, particularly their speed.

Keywords: accuracy; axioms; consensus; phylogeny; speed; supertrees; Tree of Life

1. Introduction

Although the field of supertree construction is young, tracing its origins to Gordon’s (1986) seminal contribution, there is already a rich diversity of supertree methods and variants that have been developed or outlined, some of which are in increasingly common use. Unfortunately, our understanding of these methods has not kept pace with their explosive development. The availability of multiple supertree methods means that those who seek to use them are confronted by methodological choices: which method(s) should

they use? Understanding the properties of supertree methods must be key to rational choice.

Here, we discuss some desiderata of supertree methods, the properties that we might like such methods to have. We have been inspired by the approach taken by some mathematicians (e.g., McMorris and Neumann, 1983; Barthélemy *et al.*, 1995) to the characterization and exploration of consensus methods in terms of consensus axioms, well-defined mathematical properties that they might or might not possess. Thus far, the important paper by Steel *et al.* (2000) is the only application of the axiomatic approach to supertrees. Some of the properties we discuss derive from the literature on consensus axioms, whereas others have never been discussed in that literature and originate from a biological rather than a mathematical perspective. At least some of the latter might be open to formal investigation as additional supertree (or consensus) axioms. However, we are biologists rather than mathematicians and our treatment is very informal. As biologists, we are interested in particular properties inasmuch as they impact upon our ability to do biology. Thus, we aim to clarify why the properties we discuss might be considered desirable for biologists, rather than taking their desirability to be self-evident or axiomatic.

2. What is a supertree?

A supertree is a tree that amalgamates, synthesizes, or otherwise represents the phylogenetic relationships included in a set of input trees. Under this loose definition, consensus trees are supertrees constructed in the special case of input trees with identical leaf sets. The diversity of consensus methods in evolutionary biology reflects in part a diversity of potential uses for them (Barrett *et al.*, 1991; Swofford, 1991; Wilkinson, 1994). For example, strict-consensus methods are used to summarize unanimous agreement across a set of input trees, thereby identifying those relationships that are “strictly supported” (Nixon and Carpenter, 1996). By contrast, majority-rule-consensus methods, which summarize those relationships occurring in a majority of the input trees, are used, for example, to represent the results of bootstrapping (Felsenstein, 1985, Wilkinson, 1996), jackknifing, (Farris *et al.*, 1996), quartet puzzling (Strimmer and von Haeseler, 1996), and Bayesian analyses (Larget and Simon, 1999). The utility of consensus methods depends upon what we wish of the consensus summary, and we might expect the same to be true of supertree methods.

We see four main possible uses for supertrees. For the most part, applications of supertree methods have sought to produce well-resolved, large phylogenies from sets of smaller, typically conflicting, input trees.

Here, supertrees are meta-analytic syntheses of the input trees (Mann, 1990; Sanderson *et al.*, 1998; Bininda-Emonds *et al.*, 2002) that are intended to provide a phylogenetic framework for broad comparative studies (for a review, see Gittleman *et al.*, 2004). Resolution of input-tree conflicts is hoped for, and thus “liberal-supertree” methods are used. The degree to which resolution is achieved depends upon the degree to which input trees provide differential support for conflicting relationships (as assessed by the supertree method), and also potentially by the degree of effective overlap between the input trees. Secondly, supertrees might also be used in quantitative studies of input-tree congruence. For example, outliers or unstable taxa can be identified using one or more input tree-supertree distance measures (e.g., DasGupta *et al.*, 1997) or positional congruence scores (Estabrook *et al.*, 1985). Thirdly, supertrees can be used simply to explore and identify agreement and disagreement among sets of input trees. In this case, the aim is to reveal conflict rather than to resolve it, typically through the use of “conservative-supertree” methods, with any resolution coming ultimately from additional data or new analyses sought or performed in the light of the supertree (Wilkinson *et al.*, 2001). Again, supertrees will be more or less resolved depending upon the extent of conflict and the degree of effective overlap between input trees. Finally, supertrees might be useful in identifying where limited overlap between the leaf sets of input trees is an obstacle to their amalgamation, thereby guiding further research aimed at providing effective overlap (Wilkinson *et al.*, 2001; Burleigh *et al.*, 2004). Although all these uses are important, we focus here upon liberal supertrees that are capable in principle of providing well-resolved meta-analytical syntheses in the face of conflicting input trees. Thus, we do not discuss the more conservative strict or semi-strict supertree methods (Bryant, 2002; Goloboff and Pol, 2002), which might be particularly well suited to the latter two uses.

3. Some liberal supertree methods

3.1 Matrix representations

Trees can be represented by a variety of corresponding matrices. Several supertree methods combine matrix representations of input trees into a single matrix that can be analyzed to yield a supertree. Methods differ in the form of matrix representation employed and the kind of analysis. The average consensus procedure combines pairwise-distance matrices and uses a least-squares optimality criterion in searching for the best tree (Lapointe and

Cucumel, 1997). We refer to this as a matrix representation with distances (MRD) method (Lapointe *et al.*, 2003). Most practitioners have employed matrix representations that encode trees as “pseudocharacter” data that are then analyzed with parsimony, the matrix representation with parsimony (MRP) approach to supertree construction. In standard MRP (Baum, 1992; Ragan, 1992), one binary pseudocharacter encodes each internal branch on each input tree (component or cluster coding), and standard reversible (Fitch or Wagner) parsimony is used. Irreversible MRP (Bininda-Emonds and Bryant, 1998) differs only in its use of irreversible parsimony. Purvis MRP (Purvis, 1995a) uses reversible parsimony and differs from standard MRP in the matrix representation. Each matrix element splits the members of a clade from the members of its sister group (or of all possible sister groups in the case of polytomies) and the root, with all other leaves scored as missing. In triplet and quartet MRP (Thorley, 2000; Wilkinson *et al.*, 2001), one binary pseudocharacter encodes each resolved triplet or quartet, respectively, in each input tree, and standard reversible parsimony is used. Purvis (1995a) and Rodrigo (1996) suggested, and Pisani (2002) and Ross and Rodrigo (2004) explored clique analysis as an alternative to parsimony, using component coding in their matrix representation with compatibility (MRC). In the special cases of triplet and quartet matrix representations, maximum parsimony and maximal cliques define the same optimal trees, so that MRC = MRP. A further matrix representation method involves recoding (flipping) individual entries in a component matrix representation, moving leaves into or out of clusters or from one “side” of a split to another so as to render the matrix compatible. Optimal matrix representation with flipping (MRF) supertrees are those supported by the matrices requiring the fewest recodings (Chen *et al.*, 2003; Burleigh *et al.*, 2004).

3.2 MINCUTSUPERTREE

Aho *et al.* (1981) developed a fast algorithm for amalgamating a set of compatible trees. If the trees are compatible this method returns a single supertree that contains all the input trees. Where input trees conflict, the method yields no tree. The MINCUTSUPERTREE method developed by Semple and Steel (2000) modifies the Aho *et al.* method to deal with conflicting input trees. Essentially, this is done by breaking apart conflicting clusters in a certain minimal way that ensures several desirable properties for MinCutSupertrees (for details see Semple and Steel, 2000; Page, 2002). MINCUTSUPERTREE has some Adams-consensus-like properties (Semple and Steel, 2000), and whether it is considered liberal or conservative might depend on whether the clusters in the supertree are interpreted as nestings or components (see Wilkinson, 1994).

3.3 Quartet puzzling

Quartet puzzling (Strimmer and von Haeseler, 1996) is a heuristic method for building resolved, comprehensive trees from sets of quartets that might or might not conflict. It is, therefore, a liberal supertree method. However, as normally used, it draws upon the quartet trees inferred for all possible quartets for the full set of leaves under consideration, using these in a voting procedure to determine where to add leaves to a growing tree. This is a special case, and not all quartets will be included in the input trees in the normal supertree context. Pisani and Wilkinson (2002) indicated the potential for a quartet-puzzling supertree method, but to be effective the voting procedure needs modification (Pentony *et al.*, in prep.).

In quartet puzzling, tree construction is iterated with different addition sequences and random breaking of ties. The multiple trees produced are summarized with a majority-rule consensus and the frequencies of relationships taken as an index of support (Strimmer and von Haeseler, 1996; Wilkinson *et al.*, 2003). Each quartet-puzzling iteration can be thought of as providing a fast and greedy heuristic approximation of the supertree that contains the largest number of input quartets. Thus, the method is related closely to quartet MRP. With rooted trees, triplet puzzling (i.e., quartet puzzling, but using only quartets in which one leaf is the root) would be related analogously to triplet MRC / MRP. We can envisage similar heuristics that choose a starting tree from the input trees and add taxa one at a time according to inference and fusion rules (Bryant, 1997; Dekker, 1986; Wilkinson *et al.*, 2000) and greedy local optimizations that approximate objective functions based on several tree-to-tree distances. A quartet-supertree method based on Willson's (1999, 2001) quartet methods has also been developed recently (Piaggio-Talice *et al.*, 2004).

4. Accuracy

Many consensus axioms describing desirable mathematical properties of consensus methods have been discussed, but mostly with little consideration of their relevance to what is desirable or important to biologists. In the specific context of the construction of liberal supertrees, we believe biologists are (or should be) concerned primarily with accuracy. By accuracy, we mean correspondence with actual phylogenetic relationships ("accuracy with a capital A"), rather than, for example, correspondence between the objective function of a method and heuristically selected supertrees. The ultimate aim must be to have accurate phylogenies that provide maximally useful phylogenetic frameworks for comparative biology

(Lanyon, 1993). The ability of any method to construct accurate supertrees under a range of readily modeled analytical conditions can be assessed by simulation (e.g., Bininda-Emonds and Sanderson, 2001; Chen *et al.*, 2003; Lapointe and Levasseur, 2004; Piaggio-Talice *et al.*, 2004; Ross and Rodrigo, 2004). However, the ability of methods to produce accurate trees depends very much on properties of the data, and, insights from simulations notwithstanding, we do not know how accurate real supertrees are for the most part. In the absence of an assessment of accuracy, we can examine other properties as surrogates. For example, we might investigate whether supertrees include relationships that we might reasonably expect to be present, or, conversely, relationships that we would not expect. Similarly, we can address whether the resolution of conflict is affected by properties of input trees other than those we might expect the resolution to be based upon (i.e., properties that are irrelevant to our understanding of the weight of support for particular relationships). The following is a far from exhaustive set of such properties.

4.1 Independence

Bryant (1997) gave formal definitions of two “independence” consensus axioms that relate to the insensitivity of consensus methods to the addition or pruning of input tree leaves (but which might be characterized in terms of any well-defined operation on trees). The independence (of irrelevant alternatives) axiom considers two profiles of trees. If the two profiles can be rendered identical by pruning some set of leaves from the trees in each profile, then, if we prune the same leaves from the consensus trees for each profile, the resulting consensus trees should also be identical. The second independence axiom states that, given a set of input trees from which some particular leaves are pruned, the consensus or supertree of the pruned trees might be expected to be the same as the pruned consensus or supertree of the full input trees. A consensus method that satisfies the second axiom must also satisfy the first (Barthélemy *et al.*, 1995).

It seems reasonable that extraneous information on the relationships of other (pruned) leaves should not impact upon the relationships inferred among the remaining leaves. There has been little investigation of independence axioms in the context of supertrees. In this context, input trees can logically entail relationships in combination that are not present in any single input tree, so that pruning selected leaves from the input trees could remove some entailed relationships and impact upon the supertree. In this case, the additional information is useful rather than irrelevant, and failing to obey independence axioms would not be undesirable necessarily. The following three properties might be related to the general idea of

independence. They are properties that biologists have found or might find desirable, but which have not been discussed much in the mathematical literature.

4.2 Sizeless

Suppose we wish each input tree to have equal weight. This might be reasonable if we had no basis for assigning differential weights. Purvis (1995a) provided an example showing a bias in standard MRP in cases of conflict towards relationships in larger trees, and Purvis MRP was proposed to remedy the bias. Subsequently, Ronquist (1996) showed that Purvis's coding method does not succeed in removing size bias, and suggested that this could be done by weighting the pseudocharacters from each input tree inversely with respect to their number. Bininda-Emonds and Bryant (1998) showed further that the size bias was with respect to the sizes of conflicting subtrees rather than the sizes of the input trees *per se*. Consequently, inverse weighting on tree size would not correct the size-related bias. Sanderson *et al.* (1998) summarized that no method was known that always weighted trees equally. Of course, this is true only for liberal-supertree methods, and does not hold for more conservative strict and semi-strict supertree methods. Page (2002) used a simple example to show a size bias (towards larger trees) in MINCUTSUPERTREE that led him to propose a modification to the method. The extent of size biases for different supertree methods is not well known. Because the addition and/or pruning of leaves will change size, methods that are not sizeless will not obey independence axioms.

Size bias seems like a serious problem if we want to weight trees equally. Such equal weighting might be justified by the principle of indifference (Keynes, 1920) if there is no basis for differential weighting of trees. However, the principle of indifference might also be invoked to justify equal weighting of components or of triplets. But, because larger (binary) trees include more components and more triplets, achieving equal weighting of these will entail unequal weighting of trees. Ronquist (1996) argued that the size bias of MRP methods was not unreasonable because larger trees contain more information. We are concerned with size biases in supertree methods only to the extent that these might promote inaccuracy. If large trees were more accurate than smaller trees in general, we would have reason to be unconcerned, but we do not think this is the case generally. Our concern is really that, whatever biases might exist, they should not be so severe as to prevent supertree methods from returning relationships that appear the best supported in terms of their frequency of replication in, or entailment by, the input trees and any additional information on their relative strength of support (see below).

An interesting approach to removing size biases would be to convert input trees with overlapping leaf sets to input trees with identical and full leaf sets by grafting leaves onto the input trees. There might be many ways of doing this for any given input tree, thereby defining a span of candidate supertrees for each input tree (Bryant, 2002). Fast heuristics might be used to generate a single “best” candidate from each input tree span that can then be amalgamated with (e.g., majority-rule) consensus. Semple and Steel (2002) have described a method for encoding a tree of any size with five multistate characters, and a suggestion that has yet to be explored is that such representations might be used to avoid size biases (Bininda-Emonds *et al.*, 2002)

4.3 Shapeless

Tree shape or balance (Shao and Sokal, 1990) is a characteristic of input trees that might reasonably be considered irrelevant to their evidential significance. We might therefore desire supertree methods that, in cases of conflict, do not favour relationships unduly in asymmetric or in symmetric trees. Several supertree methods are biased with respect to tree shape. For example, in cases of conflict, standard and irreversible MRP and MRF are biased towards relationships in asymmetric trees and Purvis MRP is biased towards relationships in symmetric trees (Wilkinson *et al.*, 2001, in prep.). These biases in the MRP methods appear to stem from the use of asymmetric distances or fit functions to define the optimal supertree.

Thorley and Wilkinson (2003) suggested that supertrees could be conceived of as trees that minimize the sum of the distances between the supertree and each input tree (see also Bryant, 2003). Hence, methods can differ in the distance metric (objective function) and the typically heuristic method used to approximate optimal trees. The distance between the supertree and an input tree in MRP is given by the fit (parsimony steps) of the matrix representation of the input tree to the supertree. With standard, irreversible, and Purvis MRP this distance is asymmetric: it is not equal to the fit of the matrix representation of the supertree (pruned of irrelevant leaves) to the input tree (Thorley and Wilkinson, 2003). In standard and irreversible MRP, symmetric trees have smaller distances to asymmetric trees than vice versa, and the reverse is true of Purvis MRP (Wilkinson *et al.*, in prep.). Shape bias of supertree methods has not been investigated extensively, and it is not known to what extent failure to be shapeless matters in practice. However, we find it difficult to conceive of any justification for such bias and would prefer shapeless methods if they exist.

4.4 Positionless

Wilkinson *et al.* (2001) presented a simple example that suggested that some MRP methods tend to resolve conflicts in favour of more crownward (Purvis, triplet) or basal (irreversible) positions of leaves that contribute to the conflict. Bininda-Emonds and Bryant (1998) also noted the apparent basal bias of irreversible MRP. As with tree shape, we find it difficult to conceive of justifications for such behaviour and would prefer supertree methods that have no such biases, an admittedly vaguely characterized property we term positionless (see also Cotton and Page, 2004). Very little is known about the extent to which existing supertree methods satisfy this potential desideratum, and further investigation would require a clearer conceptualization and quantification of the kind of positional relations referred to by biologists as “more basal” or “more crownward”.

4.5 Order invariance

We might expect that supertrees should be unaffected by the order in which input trees are processed (often termed neutrality) and, in the case of matrix representation methods, the order of leaves in the matrix (often termed symmetry or equality or anonymity). Neutrality and equality correspond to properties P1 and P2, respectively, of Steel *et al.* (2000). MINCUTSUPERTREE has both properties (Semple and Steel, 2000). Heuristic methods might or might not be order invariant (e.g., use of closest versus multiple random addition sequences in MRP, respectively), with greedy heuristics tending to sacrifice this desideratum for speed. Order invariance is desirable because we expect one accurate tree. However, the extent to which relationships in supertrees actually vary with input tree or leaf order can be determined, and could provide useful information on those relationships that are supported robustly by the input trees and those that are not.

4.6 Uniqueness

Methods that have the property of uniqueness always return a single supertree. Desiring a unique supertree might be seen as a natural consequence of desiring complete accuracy (on the assumption of only one true supertree; see Ross and Rodrigo, 2004). However, there might be good reason to prefer a method to return multiple trees (see Lapointe and Cucumel, 2002), such as when there are equally optimal solutions. MINCUTSUPERTREE and the quartet-supertree methods of Piaggio-Talice *et al.* (2004) are the only liberal-supertree methods that will always return a single tree. Uniqueness can be imposed additionally on other methods by

conjoining them to consensus methods with this property (Steel *et al.*, 2000), resulting in unique consensus supertrees with properties determined by both the supertree and the consensus methods used. With quartet puzzling, use of the majority-rule consensus to summarize the individual supertrees produced by each iteration of the method is integral to that approach to supertree construction.

4.7 Plenary

A plenary supertree is one that includes all the leaves of the input trees. Desiring a plenary supertree is a natural consequence of desiring complete accuracy. All the supertree methods that we are considering are plenary, but those methods that return multiple trees can be rendered non-plenary through the use of non-plenary consensus methods, such as reduced consensus (Wilkinson and Thorley, 2003) and agreement subtrees (Finden and Gordon, 1985; Bryant, 1997). Non-plenary supertree methods might be most useful for identifying unstable leaves, localizing conflict, and identifying areas with ineffective overlap. The plenary axiom corresponds to property P4 of Steel *et al.* (2000).

4.8 Pareto

Several important consensus axioms pertain to the extent to which relationships in the consensus are present in the input trees and vice versa. A consensus is Pareto with respect to a particular kind of relationship (e.g., clusters, nestings, triplets, among others) if all such relationships that are present in every input tree are present in the consensus. This is a very reasonable expectation if agreement is taken as strong surrogate for, and evidence of, accuracy. We therefore desire Pareto supertree methods. Most supertree methods do appear to be Pareto on one or more type of relationship. Thorley (2000) noted that the various MRP methods are Pareto on clusters (full splits, components); Chen *et al.* (2003) and Semple and Steel (2000) showed that MRF and MINCUTSUPERTREE, respectively, share this property, and we suggest this is true of MRC methods also (see Pisani, 2002). MINCUTSUPERTREE is also Pareto on nestings and triplets (Semple and Steel, 2000). We conjecture that MRC and MRP methods are Pareto with respect to the type of relationship encoded in the matrix representations (i.e., full or partial splits), but not Pareto on less-inclusive relationships. Thus, standard MRP is not Pareto on triplets (Thorley, 2000; Bininda-Emonds *et al.*, 2002; Wilkinson *et al.*, in prep.). Steel *et al.*'s (2000) properties P6 and P6' correspond to being Pareto on quartets and triplets, respectively.

In the supertree context, input trees can have different leaf sets so that it might be impossible for any relationship to be present in all input trees. We might therefore expect that all relationships in one or more input trees that are uncontradicted by other input trees would be in the supertree. However, Steel *et al.* (2000) have shown that no supertree method can have this property for rooted trees (their P7), and we expect this to be true also of unrooted trees. This is because, although a given relationship X might be uncontradicted by any input tree, collections of input trees can entail relationships contradicting X. Thus, we could weaken the condition further to expect all relationships in an input tree that are not contradicted by other input trees, singly or in combination, to be included in, or not contradicted by, the supertree. The semi-strict supertree method of Goloboff and Pol (2002) is a heuristic approach intended to satisfy this desideratum. It is unclear to what extent liberal, conflict-busting supertree methods do so, however.

4.9 Co-Pareto

A consensus is co-Pareto with respect to a particular kind of relationship if every relationship of that kind that is present in the consensus tree is present in one or more input trees. Consensus methods that do not obey this axiom are problematic if we consider that relationships that do not occur in any input tree are unsupported. In general, it is not reasonable to expect supertree methods to be co-Pareto because they might reasonably contain relationships that are entailed by the input trees in combination, but are not present in any of them singly. However, standard, irreversible, and Purvis MRP are not co-Pareto on clusters or on triplets even in the special (consensus) case of input trees with identical leaf sets (Wilkinson *et al.*, in prep.), where this requirement is reasonable.

If a method is co-Pareto, it ensures that any given relationship in the supertree is contained (displayed, included) in at least one input tree, and, therefore, that the relationship is compatible with at least one input tree. A weaker requirement is that each relationship is compatible with at least one input tree. A still weaker requirement is that no relationship in the supertree should be contradicted by all the input trees (in which case it also cannot be present in or entailed by any of them), or at least those input trees with leaf sets that make contradiction a logical possibility. We know that standard, irreversible, and Purvis MRP supertrees do not, and that MRC supertrees do satisfy this weakened co-Pareto axiom with respect to the kind of relationship encoded in the matrix representation and all relationships of lower cardinality (Wilkinson *et al.*, in prep.).

On the one hand, if a liberal supertree is conceived of as some sort of average or representation of central tendency of the input trees (Lapointe and Cucumel, 1997), then all contradicting relationships might be an acceptable compromise between conflicting relationships in the input trees. On the other hand, there would seem to be no good reason for a supertree to include any relationship that is contradicted by all the input trees because there is no obvious evidence for that relationship and clear counterevidence. We are concerned that relationships that contradict all the input trees are not likely to be accurate. Given that accuracy is our ultimate aim, we prefer supertrees that obey the weakened co-Pareto axiom. We consider methods that resolve conflict in favour of the best-supported alternatives present or entailed by the input trees as more likely to be accurate.

4.10 Weightable

We might have reason to consider some input trees, or some relationships in some input trees as better supported than others. Indeed Purvis (1995b:406) considered that “Because different kinds of source tree differ in their likelihood of being right, equal weighting of source trees cannot be defended (Barrett *et al.*, 1991).” An obvious and important desideratum for supertree methods is their capacity to use information on relative support for, or quality of, a given hypothesis so that this information can play its part in resolving conflicts in the input trees. Weighting of input trees, or of particular relationships of input trees, can be achieved to some degree by all methods. The simple expedient of replicating input trees allows all liberal methods to use differential tree weights. MRD accommodates information on support provided by branch lengths in input trees (Lapointe and Cucumel, 1997). Matrix-representation methods using discrete pseudocharacters are amenable to differential weighting of the type of relationship encoded in the matrix representation. Thus, weighting schemes can be used that reflect measures of support for components or triplets, such as bootstrap proportions (Felsenstein, 1985; Wilkinson, 1996) or decay indices (Bremer, 1988; Wilkinson *et al.*, 2000; Donoghue *et al.*, 1992). Ronquist (1996) argued the virtues of differential weighting in the context of MRP, and simulations suggest that weighting in this context can improve accuracy (Bininda-Emonds and Sanderson, 2001).

4.11 Assessable

In phylogenetic analysis, support is a much-used surrogate for accuracy. We have more faith in well-supported relationships and we endeavour to provide indices of support for relationships in phylogenetic trees. That the support

for relationships in phylogenetic supertrees should be analogously assessable is another obvious desideratum. The quartet-puzzling approach to supertree construction provides support indices directly from the variance in output trees arising from random variation in the choice of starting tree, order of addition of leaves, and the breaking of ties. MRD yields supertrees in which branch lengths reflect relative support (Lapointe *et al.*, 1994; Lapointe and Cucumel, 1997; Wilkinson *et al.*, 2003).

All supertree methods can be investigated with one or more methods designed to yield indices of support, and this is to be encouraged. For example, some authors have reported decay indices (Bremer support) for clades in MRP supertrees (e.g., Bininda-Emonds *et al.*, 1999; Liu *et al.*, 2001; Pisani *et al.*, 2002), although the utility of this particular support index has been questioned (Pisani *et al.*, 2002). Methods such as the bootstrap and jackknife could be used with all supertree methods, but a question arises as to what should be resampled. It seems natural that we would resample input trees, as suggested by Lapointe and Cucumel (2003) for assessing consensus trees. It is possible also to resample pseudocharacters in matrix representations. The latter approach was used by Purvis (1995b) who nonetheless noted that pseudocharacters derived from a single input tree are not independent. An additional potential problem with bootstrapping pseudocharacters rather than trees is that those input trees that yield more pseudocharacters (e.g., because of their large size) will tend to contribute disproportionately to the resampled data.

There are many techniques that are used to evaluate phylogenetic hypotheses inferred from primary data. We can expect that analogues of some of these will be used increasingly as the field of supertree construction matures. With methods that employ objective functions, we envisage the development and use of randomization tests of the null hypothesis that the fit of the input trees to the supertree is no better than expected by chance alone, (i.e., from randomly permuted trees; Creevey *et al.*, submitted). Rejecting the null hypothesis would be a minimum requirement for supertrees to be taken seriously. Randomization tests might also be used to identify significant outliers within sets of input trees (Lapointe and Cucumel, 2003; see also Daubin *et al.*, 2002). Additional assessment of supertrees might be attained using multiple supertree methods, particularly if we have no good basis for choosing between the alternative methods, on the basis that disagreement is suggestive of weakly supported inferences. Purvis and Webster (1999) compared standard and Purvis MRP, and found that the methods tend to agree, but that they disagree more as conflict in the input trees increases. Several workers have explored a range of weighting regimes to explore the robustness of real supertrees (e.g., Purvis, 1995b; Bininda-Emonds *et al.*, 1999; Lapointe and Kirsch, 2001; Liu *et al.*, 2001).

Alternative weighting schemes that reverse known or suspected biases might be particularly useful when methods that are known to be biased are used.

5. Practicality

To be at all useful, a supertree method must be practical. Generally, methods are used only when the major steps are implemented in software. There is considerable variation in the ease of implementing any of the methods at the present time, and we expect progress to be sufficiently rapid as to ensure that any discussion will be out of date. Thus, we do not discuss the implementation of methods here save to repeat a previous warning that applicability in practice should not be confused with acceptability in principle (Wilkinson *et al.*, 2001).

5.1 Speed

It has been stressed that supertree methods avoid difficulties of combining different data types (Sanderson *et al.*, 1998; Bininda-Emonds *et al.*, 2002), giving them a clear advantage over the alternative pathway to large trees, namely the phylogenetic analysis of combined data. They can also offer advantages in speed. It might be faster to assemble sets of input trees than to combine data, in which case this initial speed advantage is shared by all supertree methods. MINCUTSUPERTREE is a polynomial-time algorithm (property P5 of Steel *et al.*, 2000) that fulfils our desire for speedy analyses. By contrast, computational complexity increases exponentially with the number of leaves for all matrix representation methods. This necessitates the use of heuristics, and even the MINCUTSUPERTREE method can be conceived of as a heuristic for finding supertrees that minimize the sum of triplet distances to the input trees, and thus closely related to triplet MRP / MRC. Note that methods that rely on heuristics to approximate the best supertrees under some objective function need not satisfy desiderata satisfied by the exact method (Steel *et al.*, 2000). Given that most matrix representation methods use the same approaches and programs that are used in analyses of combined data, they would appear to offer no clear benefit in terms of speed over combined analyses of data.

5.2 Generality

A reasonable desideratum of all methods is that they not be restricted to special cases, particularly if those special cases are not often encountered in practice. This is not to say that a less general method would not satisfy other

desiderata that make it the method of choice in a specific context. Several supertree methods have been developed for the special case of input trees that do not conflict and cannot be applied to the more usual case of conflicting input trees (e.g., Aho *et al.*, 1981; Gordon, 1986; Steel, 1992; Thorley and Wilkinson, 2003).

Steel *et al.* (2000) have shown that some combinations of desirable properties of supertree methods can be satisfied only in the special case of rooted trees (or of trees sharing some other leaf in common). Despite major theoretical limitations, if some or all of the potential input trees are unrooted, we would still like to have methods capable of exploiting this potential. Purvis MRP cannot be applied to unrooted trees because the matrix representations encode sister-group relationships and these are defined only in rooted trees. Similarly restricted to rooted trees are methods where the objective function can be interpreted as the sum of the triplet distances between the input trees and the supertree. This includes the MINCUTSUPERTREE method, which, like the Adams consensus, is defined only for rooted trees.

Quartet methods and MRD are more general in being applicable in principle to both rooted and unrooted trees. Path-length distance matrices usually represent unrooted trees, but are applicable equally to rooted trees and are invariant with respect to different rootings. By contrast, ultrametric-distance matrices are always associated with rooted trees. Although most supertree construction has been done using rooted input trees and standard MRP, this method is also more general. In practice, matrix representations can be constructed for unrooted trees and combined with each other, alone or with matrix representations of rooted input trees. Unlike rooted trees, there are multiple equivalent matrix representations because, in the absence of a root with a fixed (but arbitrary) pseudocharacter state code, the assignment of pseudocharacter states to the subsets defined by the split is arbitrary and can be reversed with no loss of meaning. Use of any one matrix is arbitrary. This is unimportant in the case of standard (and quartet) MRP, but the results with irreversible parsimony will depend on the arbitrary choice of matrix representation, and we consider this undesirable in principle.

6. Discussion

The field of supertree construction is still young and would benefit from further discussion and clarification of what is expected of good liberal-supertree methods and the extent to which these expectations can be satisfied. This would be a useful prelude to the identification or development of good supertree methods that can be used in practice. Here, we have

discussed a few properties that might be considered supertree desiderata and highlighted some examples of methods that we know or conjecture do not satisfy these desiderata. We remain largely uncertain to what extent any failure to display these properties is important in practice, something that can be addressed through empirical or simulation studies. In addition, the important work of Steel *et al.* (2000) notwithstanding, the tasks of determining thoroughly which existing methods have these or other properties, of determining the compatibility of different desiderata, and of designing novel supertree methods with particular desirable properties await.

Acknowledgements

We thank David Gower, Bill Day, and Olaf Bininda-Emonds for helpful comments on the manuscript and the many colleagues who have contributed through discussion to the development of these ideas. This work was supported by BBSRC grant 40/G18385 to MW and by NSERC grant OGP0155251 to FJL. DP was at the Department of Biology, the Pennsylvania State University while this work was completed.

References

- AHO, A. V., SAGIV, Y., SZYMANSKI, T. G., AND ULLMAN, J. D. 1981. Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. *SIAM Journal of Computing* 10:405–421.
- BARRETT, M., DONOGHUE, M. J., AND SOBER, E. 1991. Against consensus. *Systematic Zoology* 40:486–493.
- BARTHÉLEMY, J.-P., McMORRIS, F. R., AND POWERS, R. C. 1995. Stability conditions for consensus functions defined on n -trees. *Mathematical Computer Modeling* 22:79–87.
- BAUM, B. R. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon* 41:3–10.
- BININDA-EMONDS, O. R. P. AND BRYANT, H. N. 1998. Properties of matrix representation with parsimony analyses. *Systematic Biology* 47:497–508.
- BININDA-EMONDS, O. R. P., GITTLEMAN, J. L., AND PURVIS, A. 1999. Building large trees by combining phylogenetic information: a complete phylogeny of the extant Carnivora (Mammalia). *Biological Reviews* 74:143–175.
- BININDA-EMONDS, O. R. P., GITTLEMAN, J. L., AND STEEL, M. A. 2002. The (super)tree of life: procedures, problems and prospects. *Annual Review of Ecology and Systematics* 33:265–289.
- BININDA-EMONDS, O. R. P. AND SANDERSON, M. J. 2001. An assessment of the accuracy of MRP supertree construction. *Systematic Biology* 50:565–579.
- BREMER, K. 1988. The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. *Evolution* 42:795–803.
- BRYANT, D. 1997. *Building Trees, Hunting for Trees and Comparing Trees*. Ph.D. dissertation, University of Canterbury, New Zealand.

- BRYANT, D. 2002. *Strict Consensus Supertrees*. Technical Report, School of Computer Science, McGill University, Canada.
- BRYANT, D. 2003. A classification of consensus methods for phylogenetics. In M. Janowitz, F.-J. Lapointe, F. R. McMorris, B. Mirkin, and F. S. Roberts (eds), *Bioconsensus*, pp. 163–184. American Mathematical Society, Providence, Rhode Island.
- BURLEIGH, J. G., EULENSTEIN, O., FERNÁNDEZ-BACA, D., AND SANDERSON, M. J. 2004. MRF supertrees. In O. R. P. Bininda-Emonds (ed). *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 65–85. Kluwer Academic, Dordrecht, the Netherlands.
- CHEN, D., DIAO, L., EULENSTEIN, O., FERNÁNDEZ-BACA, D., AND SANDERSON, M. J. 2003. Flipping: a supertree construction method. In M. Janowitz, F.-J. Lapointe, F. R. McMorris, B. Mirkin, and F. S. Roberts (eds), *Bioconsensus*, pp. 135–160. American Mathematical Society, Providence, Rhode Island.
- COTTON, J. A. AND PAGE, R. D. M. 2004. Tangled trees from molecular markers: reconciling conflict between phylogenies to build molecular supertrees. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 107–125. Kluwer Academic, Dordrecht, the Netherlands.
- DASGUPTA, B., HE, X., JIANG, T., LI, M., TROMP, J., AND ZHANG, L. 1997. On distances between phylogenetic trees. In M. Saks (ed.), *Proceedings of the 8th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 427–436. Association for Computing Machinery, New York.
- DAUBIN, V., GOUY, M., AND PERRIERE, G. 2002. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Research* 12:1080–1090.
- DEKKER, M. C. H. 1986. *Reconstruction Methods for Derivation Trees*. Master's thesis, Department of Mathematics and Computer Science, Vrije Universiteit, Amsterdam.
- DONOGHUE, M. J., OLSTEAD, R. G., SMITH, J. F., AND PALMER, J. D. 1992. Phylogenetic relationships of Dipsacales based on *rbcL* sequences. *Annals of the Missouri Botanical Gardens* 79:333–345.
- ESTABROOK, G. F., McMORRIS, F. R., AND MEACHAM, C. A. 1985. Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. *Systematic Zoology* 34:193–200.
- FARRIS, J. S., ALBERT, V. A., KÄLLERSJÖ, M., LIPSCOMB, D., AND KLUGE, A. G. 1996. Parsimony jackknifing outperforms neighbor-joining. *Cladistics* 12:99–124.
- FELSENSTEIN, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39:783–791.
- FINDEN, C. R. AND GORDON, A. D. 1985. Obtaining common pruned trees. *Journal of Classification* 2:225–276.
- GITTLEMAN, J. L., JONES, K. E., AND PRICE, S. A. 2004. Supertrees: using complete phylogenies in comparative biology. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 439–460. Kluwer Academic, Dordrecht, the Netherlands.
- GOLOBOFF, P. A. AND POL, D. 2002. Semi-strict supertrees. *Cladistics* 18:514–525.
- GORDON, A. D. 1986. Consensus supertrees: the synthesis of rooted trees containing overlapping sets of labeled leaves. *Journal of Classification* 3:31–39.
- KEYNES, J. M. 1920. *A Treatise on Probability*. MacMillan, London.
- LANYON, S. M. 1993. Phylogenetic frameworks: towards a firmer foundation for the comparative approach. *Biological Journal of the Linnean Society* 49:45–61.

- LAPOINTE, F.-J. AND CUCUMEL, G. 1997. The average consensus procedure: combination of weighted trees containing identical or overlapping sets of taxa. *Systematic Biology* 46:306–312.
- LAPOINTE, F.-J. AND CUCUMEL, G. 2002. Multiple consensus trees. In K. Jajuga, A. Sokolowski, and H.-H. Bock (eds), *Classification, Clustering and Data Analysis: Recent Advances and Applications*, pp. 359–364. Springer-Verlag, Berlin.
- LAPOINTE, F.-J. AND CUCUMEL, G. 2003. How good can a consensus get? Assessing the reliability of consensus trees in phylogenetic studies. In M. Janowitz, F.-J. Lapointe, F. R. McMorris, B. Mirkin, and F. S. Roberts (eds), *Bioconsensus*, pp. 205–220. American Mathematical Society, Providence, Rhode Island.
- LAPOINTE, F.-J. AND KIRSCH, J. A. W. 2001. Construction and verification of a large phylogeny of marsupials. *Australian Mammalogy* 3:9–22.
- LAPOINTE, F.-J., KIRSCH, J. A. W., AND BLEIWEISS, R. 1994. Jackknifing of weighted trees: validation of phylogenies reconstructed from distance matrices. *Molecular Phylogenetics and Evolution* 3:256–267.
- LAPOINTE, F.-J. AND LEVASSEUR, C. 2004. Everything you always wanted to know about the average consensus, and more. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 87–105. Kluwer Academic, Dordrecht, the Netherlands.
- LAPOINTE, F.-J., WILKINSON, M., AND BRYANT, D. 2003. Matrix representations with parsimony or with distances: two sides of the same coin? *Systematic Biology* 52:865–868.
- LARGET, B. AND SIMON, D. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution* 16:750–759.
- LIU, F.-G. R., MIYAMOTO, M. M., FREIRE, N. P., ONG, P. Q., TENNANT, M. R., YOUNG, T. S., AND GUGEL, K. F. 2001. Molecular and morphological supertrees for eutherian (placental) mammals. *Science* 291:1786–1789.
- MANN, C. 1990. Meta-analysis in the breech. *Science* 249:476–479.
- MCMORRIS, F. R. AND NEUMANN, D. 1983. Consensus functions defined on trees. *Mathematical Social Sciences* 4:131–136.
- NIXON, K. C. AND CARPENTER, J. M. 1996. On consensus, collapsibility and clade concordance. *Cladistics* 12:305–201.
- PAGE, R. D. M. 2002. Modified mincut supertrees. In R. Guigó and D. Gusfield (eds), *Algorithms in Bioinformatics, Second International Workshop, WABI 2002, Rome, Italy, September 17–21, 2002, Proceedings*, pp. 537–552. Springer, Berlin.
- PIAGGIO-TALICE, R., BURLEIGH, J. G., AND EULENSTEIN, O. 2004. Quartet supertrees. In O. R. P. Bininda-Emonds (ed). *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 173–191. Kluwer Academic, Dordrecht, the Netherlands.
- PISANI, D. 2002. *Comparing and Combining Data and Trees in Phylogenetic Analysis*. Ph.D. dissertation, Department of Earth Sciences, University of Bristol, United Kingdom.
- PISANI, D. AND WILKINSON, M. 2002. MRP, taxonomic congruence and total evidence. *Systematic Biology* 51:151–155.
- PISANI, D., YATES, A. M., LANGER, M. C., AND BENTON, M. J. 2002. A genus-level supertree of the Dinosauria. *Proceedings of the Royal Society of London B* 269:915–921.
- PURVIS, A. 1995a. A modification to Baum and Ragan's method for combining phylogenetic trees. *Systematic Biology* 44:251–255.
- PURVIS, A. 1995b. A composite estimate of primate phylogeny. *Philosophical Transactions of the Royal Society of London B* 348:405–421.

- PURVIS, A. AND WEBSTER, A. J. 1999. Phylogenetically independent comparisons and primate phylogeny. In P. C. Lee (ed.), *Comparative Primate Socioecology*, pp. 44–70. Cambridge University Press, Cambridge.
- RAGAN, M. A. 1992. Phylogenetic inference based on matrix representation of trees. *Molecular Phylogenetics and Evolution* 1:53–58.
- RODRIGO, A. G. 1996. On combining cladograms. *Taxon* 45:267–274.
- RONQUIST, F. 1996. Matrix representation of trees, redundancy, and weighting. *Systematic Biology* 45:247–253.
- ROSS, H. A. AND RODRIGO, A. G. 2004. An assessment of matrix representation with compatibility in supertree construction. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 35–63. Kluwer Academic, Dordrecht, the Netherlands.
- SANDERSON, M. J., PURVIS, A., AND HENZE, C. 1998. Phylogenetic supertrees: assembling the trees of life. *Trends in Ecology and Evolution* 13:105–109.
- SEMPLE, C. AND STEEL, M. 2000. A supertree method for rooted trees. *Discrete Applied Mathematics* 105:147–158.
- SEMPLE, C. AND STEEL, M. 2002. Tree reconstruction from multistate characters. *Advances in Applied Mathematics* 28:169–184.
- SHAO, K. AND SOKAL, R. R. 1990. Tree balance. *Systematic Zoology* 39:266–276.
- STEEL, M., DRESS, A. W. M., AND BÖCKER, S. 2000. Simple but fundamental limitations on supertree and consensus tree methods. *Systematic Biology* 49:363–368.
- STRIMMER, K. AND VON HAESELER, A. 1996. Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. *Molecular Biology and Evolution* 13:964–969.
- SWOFFORD, D. L. 1991. When are phylogeny estimates from molecular and morphological data incongruent? In M. M. Miyamoto and J. Cracraft, (eds), *Phylogenetic Analyses of DNA Sequences*, pp. 295–333. Oxford University Press, New York.
- THORLEY, J. L. 2000. *Cladistic Information, Leaf Stability and Supertree Construction*. Ph.D. dissertation, School of Biological Sciences, University of Bristol, United Kingdom.
- THORLEY, J. L. AND WILKINSON, M. 2003. A view of supertree methods. In M. Janowitz, F.-J. Lapointe, F. R. McMorris, B. Mirkin, and F. S. Roberts (eds), *Bioconsensus*, pp. 185–193. American Mathematical Society, Providence, Rhode Island.
- WILKINSON, M. 1994. Common cladistic information and its consensus representation: reduced Adams and reduced cladistic consensus trees and profiles. *Systematic Biology* 43:343–368.
- WILKINSON, M. 1996. Majority-rule reduced consensus methods and their use in bootstrapping. *Molecular Biology and Evolution* 13:437–444.
- WILKINSON, M., LAPONTE, F.-J., AND GOWER, D. J. 2003. Branch lengths and support. *Systematic Biology* 52:127–130.
- WILKINSON, M. AND THORLEY, J. L. 2003. Reduced consensus. In M. Janowitz, F.-J. Lapointe, F. R. McMorris, B. Mirkin, and F. S. Roberts (eds), *Bioconsensus*, pp. 195–203. American Mathematical Society, Providence, Rhode Island.
- WILKINSON, M., THORLEY, J. L., LITTLEWOOD, D. T. J., AND BRAY, R. A. 2001. Towards a phylogenetic supertree for the Platyhelminthes? In D. T. J. Littlewood and R. A. Bray (eds), *Interrelationships of the Platyhelminthes*, pp. 292–301. Chapman-Hall, London.
- WILKINSON, M., THORLEY, J. L., AND UPCHURCH, P. M. 2000. A chain is no stronger than its weakest link: double decay analyses of phylogenetic hypotheses. *Systematic Biology* 49:754–776.
- WILLSON, S. J. 1999. Building phylogenetic trees from quartets by using local inconsistency measures. *Molecular Biology and Evolution* 16:685–693.

WILLSON, S. J. 2001. An error correcting map for quartets can improve the signals for phylogenetic trees. *Molecular Biology and Evolution* 18:344–351.

Chapter 11

TAXONOMY, SUPERTREES, AND THE TREE OF LIFE

Roderic D. M. Page

Abstract: Some of the main practical impediments to the application of supertrees in large-scale phylogenetic analysis are inconsistent use of taxonomic names, trees incorporating taxa of different ranks, and poor taxonomic overlap between different phylogenetic studies. This chapter considers these problems and suggests some solutions. The notion of a “classification graph” is introduced to test for consistency between higher-level classifications. One strategy for coping with poor taxonomic overlap is to use a constraint tree that specifies some taxonomic groups that must appear in the supertree.

Keywords: classification graphs; cluster graphs; constraints; MINCUTSUPERTREE, taxonomy

1. Introduction

If supertrees are to play a role in constructing the Tree of Life, then there are several issues that need to be addressed before their use becomes feasible. This chapter explores some of these issues, motivated in part by my experience while implementing the MINCUTSUPERTREE algorithm (Semple and Steel, 2000) and integrating it into the phylogenetic database TreeBASE (Piel *et al.*, 2002; <http://www.treebase.org>). I begin by rehearsing briefly the arguments for and against supertrees, and then go on to outline some problems that arise once we move beyond developing algorithms and try to deploy supertree methods on real data. Some of the ideas developed here were hinted at in my earlier work on the MINCUTSUPERTREE algorithm

(Page, 2002). In each section I outline what I think are some open problems that deserve attention.

1.1 Why use supertrees?

There are several reasons why supertrees have a role in phylogenetic inference. First, there is the issue of data combinability. Phylogenetically useful data comes in a variety of forms, including discrete characters (e.g., sequences, morphology, behaviour), distances (e.g., DNA-DNA hybridization, genomic signatures, gene composition), and signed permutations (e.g., genome order). Not all these data can be treated readily under a single framework. Discrete data can be handled using standard parsimony methods, and likelihood methods can be applied to both sequence and morphological data, but there remains a range of data that are not amenable directly to either approach. Furthermore, if we are interested in organismal phylogeny, then data might require additional transformation before being used. For example, we could argue that nucleotide and protein sequences are not directly characters of organisms, but rather are characters of genes (Slowinski and Page, 1999) such that organismal phylogenies are not obtained directly from gene sequences, but via gene trees. This notion of gene tree parsimony is discussed by Cotton and Page (2004). Under this view, even if sequence data can be combined directly, it might not be appropriate to do so.

The second argument for the use of supertrees concerns computational tractability. As is well known, inferring optimal phylogenetic trees from discrete and distance data is NP-complete. This places practical limits on the size of tree we can hope to construct and still be reasonably confident that we have an optimal or near-optimal solution. Although the phylogenetic community has made a lot of progress in developing sophisticated search algorithms, some of which scale to tens of thousands of sequences, constructing trees on the scale of the Tree of Life does not look feasible at present. One approach to this problem is to adopt a divide-and-conquer strategy; namely, we build smaller trees and assemble these into a larger tree (e.g., see Roshan *et al.*, 2004). This is the supertree approach. For this argument to be compelling, supertree algorithms must be more efficient than tree-building algorithms. The only polynomial-time supertree algorithms to date are ONETREE (also known as BUILD; Aho *et al.*, 1981) and MINCUTSUPERTREE (Semple and Steel, 2000). The most popular algorithm for supertrees, matrix representation with parsimony (MRP; Baum, 1992; Ragan, 1992; Baum and Ragan, 2004), is NP-complete, as are others such as minimum flipping (MRF; Chen *et al.* 2002, Burleigh *et al.*, 2004) and compatibility methods (see Goloboff and Pol, 2002; Ross and Rodrigo,

2004). Indeed, it seems that just about any formulation of the supertree problem where there is a global optimality criterion is NP-complete (Bryant, 1997). This emphasizes the attractiveness of a polynomial-time algorithm such as MINCUTSUPERTREE, especially if the goal is to integrate supertree construction into a phylogenetic database such as TreeBASE.

1.2 Against supertrees

Not everyone is enamored of supertrees (e.g., Novacek, 2001; see also Gatesy and Springer, 2004), and there are several good arguments that can be made against their use. Supertrees are one step removed from the underlying data, and there is considerable concern that trees being inputted into a supertree analysis lack some measure of how well supported they are. There have been various attempts to address this problem, either through weighting individual trees or through weighting nodes within trees. Neither solution is terribly attractive and doesn't address the more serious problem demonstrated by Barrett *et al.* (1991). Given two data matrices, it is possible that the optimal tree for the combined data matrix contains nodes that are not in the optimal trees for either of the two matrices when analyzed separately ("signal enhancement"). Put another way, some phylogenetic groups that receive limited support from small matrices start to emerge as well supported in larger analyses (Gatesy *et al.*, 1999; see also Bininda-Emonds *et al.* 2004).

One way to address at least some of this concern would be to analyze confidence sets of trees for each data set, rather than a single tree. For example, given a suite of k data sets, we might perform a bootstrap analysis on each data set, generating a confidence set of n trees (Sanderson, 1989) for each matrix. We would then draw a single tree from each bootstrap set, compute a supertree, and then repeat this until we have constructed n supertrees. We would then summarize the effects of uncertainty in the input trees on our supertree by computing a consensus of the n supertrees. Cotton and Page (2002) used this approach in their gene tree parsimony analysis of vertebrate phylogeny.

Given that a supertree is akin to a consensus tree (Bryant, 2003), the objections raised against consensus trees as estimates of phylogeny (Miyamoto, 1985) can be leveled at supertrees, particularly methods such as MINCUTSUPERTREE, which yields a single tree with properties related to the Adams consensus tree (Adams, 1986). However, just as with consensus trees, this argument can be mitigated somewhat by not treating the supertree as reflecting a phylogeny directly, but rather summarizing a set of phylogenies.

There are also practical issues about the independence of input trees. Phylogenetic studies often incorporate data from previous studies (e.g., a

study on bird phylogeny might incorporate a DNA sequence for the chicken that was obtained in an earlier study). As a result, supertree analysis that use trees extracted directly from the literature are likely to contain a mixture of independent and dependent studies. The extent of this problem is illustrated nicely by Gatesy *et al.* (2002) in their discussion of Liu *et al.*'s (2001) study of mammalian phylogeny (see also Bininda-Emonds *et al.*, 2004).

Perhaps the most powerful argument against using supertrees is the effect of taxon sampling on the accuracy of phylogenetic trees. Supertree methods, by definition, assemble larger trees from suites of smaller trees. If smaller trees are less accurate than larger trees for the same data, then we might expect supertree methods to perform less well than direct analyses of combined data matrices. The effect of taxon sampling on phylogenetic reconstruction is the subject of debate currently (Pollock *et al.*, 2002), but if smaller trees are consistently less accurate than larger trees, then the utility of supertree methods might be compromised.

2. Taxonomy

Michael Sanderson proposed the following supertree challenge as part of the “Deep Green Challenges” (see <http://www.life.umd.edu/labs/delwiche/deepgreen/DGchallenges.html>):

“The TreeBASE database (<http://www.herbaria.harvard.edu/treebase>) currently contains over 1000 phylogenies with over 11,000 taxa among them. Many of these trees share taxa with each other and are therefore candidates for the construction of composite phylogenies, or “supertrees”, by various algorithms. A challenging problem is the construction of the largest and “best” supertree possible from this database. “Largest” and “best” may represent conflicting goals, however, because resolution of a supertree can be easily diminished by addition of “inappropriate” trees or taxa.”

This challenge has, to my knowledge, not been met, nor even attempted. Indeed, I suggest that the task of assembling a (meaningful) supertree from the trees stored in TreeBASE is doomed to failure because of the lack of a consistent biological taxonomy, and because the taxa referred to in those trees occupy all levels in the taxonomic hierarchy.

2.1 Taxonomic consistency

When assembling a supertree from a suite of smaller trees, one of the most fundamental assumptions we are making is that nodes referred to by the

same label are the same, and nodes labeled by different labels are different. If two trees have a node labeled “A”, then both nodes represent the same entity. Conversely, the same entity is not going to be called “A” in one tree and “B” in another. This simple condition requires that the biological taxonomy of the organisms we are investigating is correct or at least consistent (see Bininda-Emonds *et al.*, 2004). The reality is that this is rarely the case. Different authors use different names for the same taxa, and databases such as TreeBASE and GenBank do not impose a consistent taxonomy rigorously. As a result, we cannot assume that all nodes labeled “A” are the same. There can be multiple names for the same taxon (e.g., synonyms or plain misspellings), and the same name can mean different things in different classifications. This is not to say that all differences between classifications are synonyms or errors. Some differences are the result of legitimate scientific disagreements. However, at some level, there has to be some consistent labeling of the entities we are interested in.

2.1.1 Problem: unique identifiers for taxa. Large-scale phylogenetic analysis that uses multiple data sources will require standardized names. This is a particular concern for phylogenetic databases. There are efforts to compile lists of taxonomic names (e.g., the Species 2000 Project), but until such lists are integrated with phylogenetic and sequence databases, the automated construction of supertrees from databases will be frustrated.

3. Higher taxa

Most of the supertree literature considers the case where the input trees all contain taxa at the same taxonomic rank (e.g., genera). However, in some cases, our input trees might include taxa at different levels in the hierarchy. This is a common occurrence in morphological studies, where a terminal taxon might represent a synthesis of information for several subordinate taxa. This is in contrast to molecular studies, where a higher taxon is represented by one or more exemplars — a sequence from a single organism that is used to represent the larger taxonomic group.

To illustrate the problem, consider the two spider trees shown in Figure 1 (from Coddington, 1991). On the face of it, we cannot create a supertree from these two trees because they have only a single terminal taxon (“Austrochilidae”) in common. However, note that “Araneoclada” is a leaf in the first tree and an internal node in the second, and “Araneomorphae” and “Neocribellatae” are internal labels in both trees. Hence, there is in fact considerable overlap between these trees. Current supertree methods do not handle such trees correctly. In some cases, we can simply substitute taxa

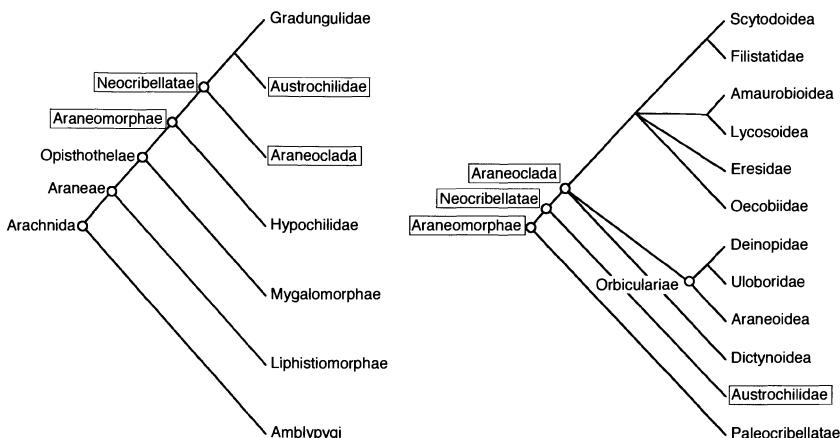


Figure 1. Two trees for spiders and related taxa. Labeled internal nodes are distinguished by \circ , and labels that are the same in the two trees are enclosed in boxes. Although both trees include the taxa Araneomorphae, Neocribellatae, Araneoclada, and Austrochilidae, the only leaf the two trees share is the Austrochilidae. Trees were obtained from study S1x6x97c14c42c30 in TreeBASE (<http://www.treebase.org>).

(Wilkinson *et al.*, 2001), but in the example shown in Figure 1 we can't simply replace the node labeled "Araneoclada" in the first tree with the subtree rooted at the node labeled "Araneoclada" in the second tree because this will not take into account the fact that the two trees also share the "" and the "Araneomorphae".

3.1 Classification graph

To accommodate trees with internal node labels (Figure 1), we need to be able to take a set of two or more input trees and decide whether the relationships among the labeled nodes are consistent. Here I sketch one possible approach that makes use of a "classification graph" G . This is a directed graph where the nodes correspond to the labeled nodes in the input trees (or "classifications"), and the edges represent the relative hierarchical ordering of the taxa with whose names the nodes are labeled.

Given a set of trees \mathcal{T} , where all leaf nodes and at least some internal nodes are labeled, we construct the graph in the following way. First, for each tree T_i in \mathcal{T} , we contract all edges with unlabelled end nodes. The resulting trees now depict just the relationships among named taxa. For each pair of nodes (x, y) in each tree T_i in \mathcal{T} , if x is the immediate ancestor of y we add an edge (y, x) between the corresponding nodes in G (unless such an edge exists already). Figure 2 shows the classification graph for the two spider trees shown in Figure 1. Based on the relationships in this graph, we

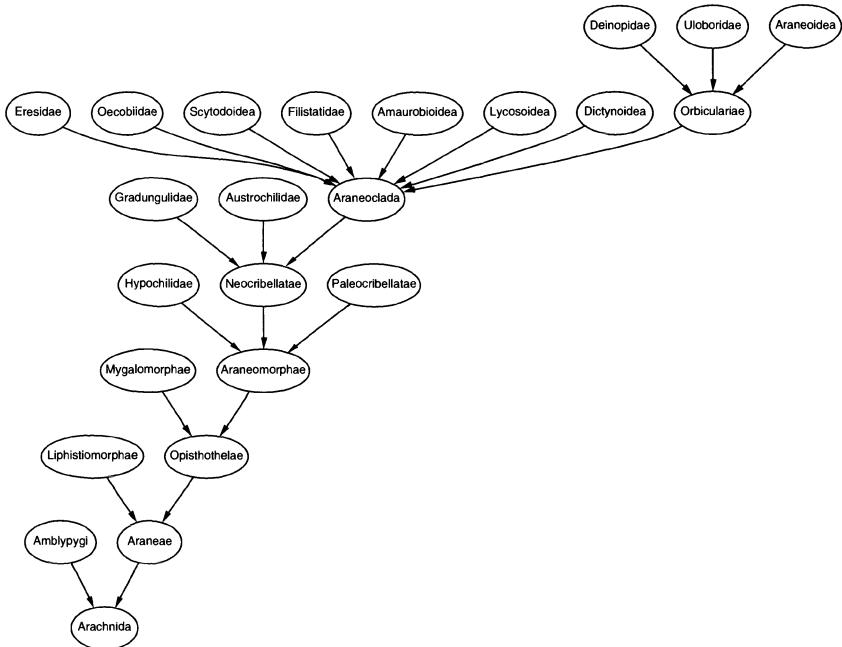


Figure 2. Classification graph for two spider trees shown in Figure 1.

can construct a supertree for the two spider trees over which both trees agree (Figure 3). This supertree is consistent with the classification graph, but is more resolved. For example, in the classification graph (Figure 2) there are eight nodes that are linked directly to the node labeled “Araneoclada”, whereas in the supertree (Figure 3), the equivalent node has only three immediate descendants.

3.2 Consistency of higher classifications

We can impose some basic requirements on a classification graph G . The first is that the graph be *connected*. If it is not, then there is no overlap in classifications and the trees can't be combined meaningfully. The second requirement is that the graph be *acyclic*. A directed cycle implies that the hierarchical order of taxon names is not consistent internally. Let $x \prec y$ mean that taxon x is higher in the taxonomic hierarchy than taxon y , so that \prec Araneoclada (see Figure 2). If one tree had \prec Araneoclada and another had Araneoclada \prec , then the two classifications would be inconsistent. Finally, there should be only one node in G with out-degree 0 (i.e., the number of edges leaving the node is zero). This node is the “root” node.

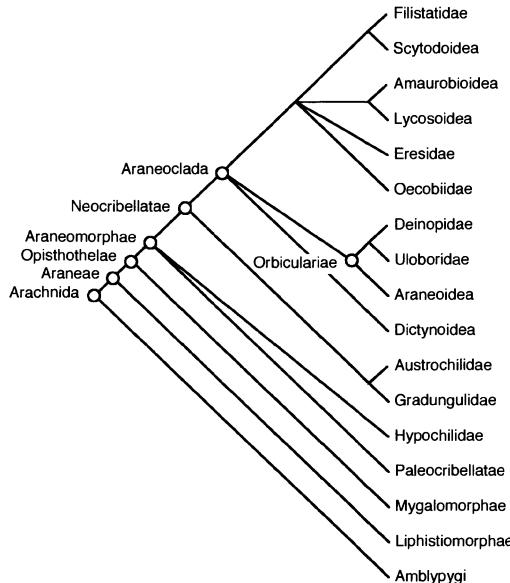


Figure 3. Supertree for the two spider trees shown in Figure 1 that is consistent with the classification graph shown in Figure 2.

The classification graph for the spiders satisfies these criteria. However, these requirements alone are not sufficient. We need to be able to represent the classifications by a single tree, and in a tree there is a unique path from every node to the root. This need not always be true in a classification graph.

3.3 Paths

Consider the three trees shown in Figure 4. The classification graph for these three trees is not a tree (Figure 5) — the nodes “Hominoidae” and “Primates” have out-degree greater than one. This is a consequence of the partial overlap among the three trees. The first tree tells us that Mammalia \prec Hominoidae, the second adds the information that Primates \prec Hominoidae, and the third contributes Mammalia \prec Primates. Taken together, these taxa are ordered Mammalia \prec Primates \prec Hominoidae.

To extract this information from the classification graph, we can do the following:

1. Visit each node n with out-degree greater than one.
2. Construct the subgraph of G that contains just the nodes and edges on the paths between node n and the root node. Figure 6 shows the subgraph for Hominoidae.

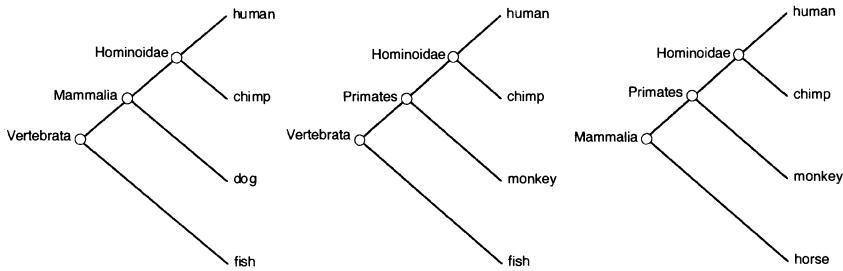


Figure 4. Three hypothetical trees for some vertebrate taxa.

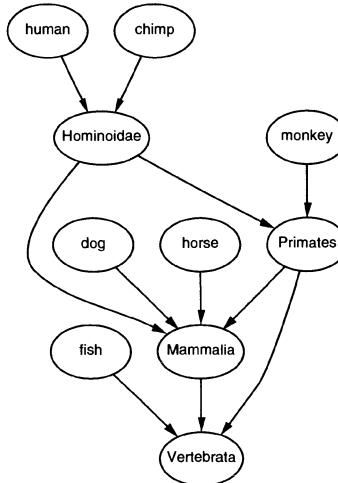


Figure 5. Classification graph for the three trees shown in Figure 4.

3. Sort the subgraph topologically so that we have a linear ordering of each node in the subgraph.
4. Check that the linear ordering is unique.

A topological sorting of a directed graph is a linear ordering such that if there is an edge from node x to node y , then x appears before y in the ordering. The order in which one puts one's clothes on when getting dressed is an example of topological sorting (Cormen *et al.*, 1990). Some items must be put on before others (e.g., socks must go on before shoes), whereas other items can go on in any order with respect to each other (e.g., pants and socks). If we construct a graph where a directed edge connects a pair of items (x, y) where x must be put on before y , then sorting the resulting graph topologically yields a order in which we can get dressed. Note that there can

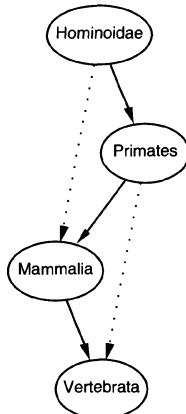


Figure 6. Subgraph of the graph shown in Figure 5 for the Hominoidae. Edges that link nodes that are not adjacent in the topological sort order of the graph are drawn with dotted lines.

be more than one possible ordering — there is more than one way to get dressed.

For the classification graph, we need there to be a single ordering. We check for this by testing whether there is an edge between each adjacent member of the ordering. If this is the case, then there is a single ordering of the subgraph, and we can extract an unambiguous tree from the classification graph by deleting all edges between nonadjacent nodes. For the Hominoidae subgraph (Figure 6), the topological order of the nodes is Hominoidae, Primates, Mammalia, Vertebrata. The edges (Hominoidae, Mammalia) and (Primates, Vertebrata) can be deleted. If there is more than one ordering, then the relationship between the classifications is ambiguous, in which case it is not clear how they can be combined.

3.3.1 Problem: classification graphs. The method above for testing classification consistency is an outline only, and needs further development. For instance, it uses only information contained within the trees. However, classifications that embody ranks (such as the Linnaean system) provide additional information. Given that Primates is an order, and Mammalia is a class, we know that $\text{Primates} \prec \text{Mammalia}$ is impossible logically. This information could be added to the classification graph.

3.3.2 Problem: consistency of trees with labeled internal nodes. We know that for rooted, terminally labeled trees the question of whether they are consistent can be answered efficiently (Aho *et al.*, 1981; Steel, 1992), but is there an efficient way to test for consistency of a suite of trees where both

the internal and terminal nodes are labeled? (This problem was solved subsequently to it being posed by me by Daniel and Semple (2004).).

4. Taxonomic overlap

A major problem facing the practical use of supertree methods is the poor degree of taxonomic overlap between trees derived from different sources. The degree of taxonomic overlap can be visualized using “cluster graphs” (Sanderson *et al.*, 1998), where the nodes represent individual trees. Two nodes, x and y , are connected by an edge if, for some fixed k , the corresponding trees have at least k leaves in common. At a minimum, we need two leaves in common to construct a supertree. If we construct a cluster graph for $k = 2$ and that graph is not connected, then we cannot construct a supertree for all the trees of interest.

4.1 Cluster graphs

We can explore the degree of taxonomic overlap in a set of input trees by constructing cluster graphs for different values of k and finding the components of these graphs. For real data, the results can be disappointing. Figure 7 shows an example for birds based on 143 generic-level phylogenies for birds assembled from the literature (Lauk, 2002). If we require minimal overlap (i.e., $k = 2$), almost all (129) of the phylogenies form a single component, with a few isolated studies remaining. As we increase the amount of minimum overlap required, the major component gets progressively smaller, until $k = 5$, when the set of 143 bird trees fragments into two components with 27 and 29 trees, respectively, and numerous smaller components. Given that the degree of taxonomic overlap between input trees is a key predictor of the accuracy of the derived supertree (Bininda-Emonds and Sanderson, 2001; Chen *et al.*, 2002), then the greater the degree of overlap the better the resulting supertree is likely to be. For the bird example, this comes at the price of being able to construct supertrees for a limited number of taxa only using subsets of the original sets of trees. For $k = 5$, over a third of the input trees do not overlap with any other tree and hence have to be discarded.

4.1.1 Problem: cluster graphs for higher taxa. The trees shown in Figure 1 have only a single leaf in common, and hence have $k = 1$, despite the large degree of taxonomic overlap. Can we modify the construction of cluster graphs to handle internally labeled nodes?

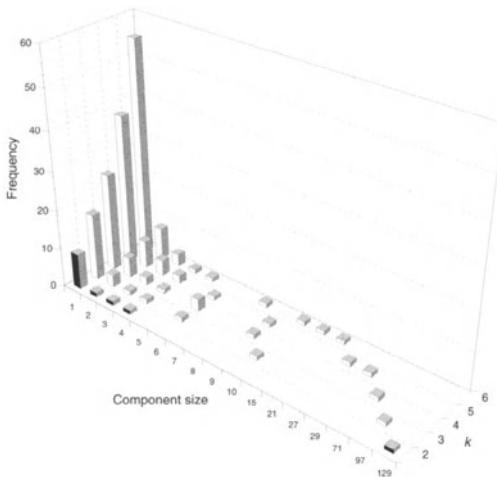


Figure 7. Size distributions of components of cluster graphs for 143 genus-level bird phylogenies constructed for different levels of overlap (data from Lauk, 2002).

4.2 Bicliques

Burleigh *et al.* (2004) describe the use of bicliques as a tool for identifying sets of trees that can be combined in a supertree analysis. Given a bipartite graph where one set of nodes represents trees, and the other set represents taxa, we can draw an edge between a tree node and a taxon node if that tree contains that taxon (Figure 8). A subgraph where every tree node is connected to every taxon node is a biclique (for details, see Burleigh *et al.*, 2004).

I used an implementation of the Alexe *et al.* (2000) biclique algorithm, kindly made available by Oliver Eulensteiner, to investigate the 143-tree bird data set. This program found 658 maximal bicliques for these trees. At one extreme, there is a biclique comprising a single tree and 67 taxa, which corresponds to the largest source tree in the data set (from Chu, 1995), and at the other extreme there is a biclique comprising 32 trees and a single taxon — the most common bird genus among the source trees, the chicken *Gallus*.

Although appealing, bicliques can be problematic to interpret. Each biclique has (m, n) nodes comprising m trees and n taxa. A biclique of three trees and ten taxa, $(m, n) = (3, 10)$, has the same number of nodes (and edges) as a biclique of ten trees and three taxa, $(m, n) = (10, 3)$. However, these two bicliques can have different implications for a supertree analysis. The first identifies a small set of trees with a reasonable number of taxa in common, whereas the second biclique comprises more trees but with fewer shared taxa.

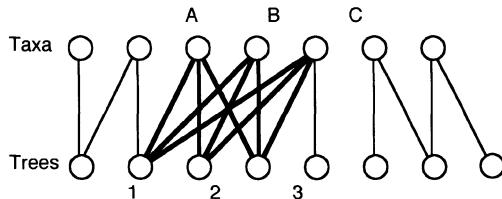


Figure 8. A bipartite graph with a maximal biclique (bold lines) comprising taxa A, B, C, and trees 1, 2, 3 (from Burleigh *et al.*, 2004).

The minimum required overlap for supertree construction is two taxa, so we are interested in bicliques where at least $n = 2$. For the bird trees, the two largest such bicliques contain only ten trees (i.e., $(m, n) = (10, 2)$). The two taxa shared by all trees in the first biclique are the chicken (*Gallus*) and the duck (*Anas*), the second biclique comprises the chicken and the ostrich (*Struthio*). Note that bicliques require that all trees share the *same* two taxa, whereas the cluster graph approach for $k = 2$ requires merely that each pair of trees shares any two taxa. Consequently, for an overlap of two taxa we obtain a cluster graph with a 129-tree component, but a biclique of only ten trees.

The taxa found in the two $(10, 2)$ bicliques are all farmed or domesticated, which accounts partly for their frequency in the set of trees. It is not only the degree of overlap between trees, but also the relationships among the overlapping taxa that influence the efficacy of supertree reconstruction (Wilkinson *et al.*, 2001). Chickens and ducks are relatively closely related, and the ostrich is a member of the putatively basal bird clade, the ratites. Given this, trees sharing these taxa place very few constraints on relationships among other bird genera (which comprise the bulk of modern birds). Hence, even the trees identified by the bicliques are unlikely to be very informative about avian relationships.

4.3 Higher taxa

In some cases, the lack of taxonomic overlap is due to studies using different “exemplars” of higher taxa. For example, if one study on mammalian phylogeny used a mouse to represent rodents, and another study used a rat, then a supertree for those two studies need not group mouse and rat together. One solution would be to impose a constraint on the set of possible solutions (e.g., by specifying a constraint tree; Constantinescu and Sankoff, 1986) that must be displayed by the supertree. In the previous example, we could require that all solutions grouped mouse and rat together.

Imposing constraints in MRP is straightforward because software such as PAUP* (Swofford, 2002) implements constraint trees. We need an equivalent procedure for MINCUTSUPERTREE. Before describing an approach to imposing constraints, I will describe briefly the original MINCUTSUPERTREE algorithm (Semple and Steel, 2000).

4.4 MINCUTSUPERTREE

Semple and Steel's (2000) MINCUTSUPERTREE algorithm takes as input a set of k rooted trees \mathcal{T} and a set of species $S = \bigcup_{i=1}^k \mathcal{L}(T_i) = \{x_1, \dots, x_n\}$, where $\mathcal{L}(T_i)$ is the set of leaves in tree T_i . The algorithm constructs the graph $S_{\mathcal{T}}$ recursively. The nodes in $S_{\mathcal{T}}$ are terminal taxa, and nodes a and b are connected if a and b are in a proper cluster in at least one of the input trees (i.e., if there is a tree in which the most recent common ancestor of a and b is not the root of that tree). The algorithm proceeds as follows:

procedure MINCUTSUPERTREE(\mathcal{T})

1. **if** $n = 1$, **then return** a single node labeled by x_1 .
2. **if** $n = 2$, **then return** a tree with two leaves labeled by x_1 and x_2 .
3. Otherwise, construct $S_{\mathcal{T}}$ as described above.
4. **if** $S_{\mathcal{T}}$ is disconnected **then**
 - Let S_i be the components of $S_{\mathcal{T}}$.
 - else** Create graph $S_{\mathcal{T}} / E^{\max}_{\mathcal{T}}$ and delete all edges in $S_{\mathcal{T}} / E^{\max}_{\mathcal{T}}$ that are in a minimum cut set of $S_{\mathcal{T}}$. Let S_i be the resulting components of $S_{\mathcal{T}} / E^{\max}_{\mathcal{T}}$.
5. **for** each component S_i , **do**
 - $T_i = \text{MINCUTSUPERTREE}(\mathcal{T} | S_i)$, where $\mathcal{T} | S_i$ is the set of input trees with all species not in S_i pruned.
6. Construct a new tree \mathcal{T} by connecting the roots of the trees T_i to a new root r .
7. **return** T

end

The key difference between the ONETREE algorithm and MINCUTSUPERTREE lies in step 4. In ONETREE, if the graph $S_{\mathcal{T}}$ is connected (i.e., comprises a single component) then the algorithm exits, returning the result that the input trees are not consistent. Semple and Steel modified ONETREE by ensuring that $S_{\mathcal{T}}$ yields more than one component by using minimum cuts. Given a connected graph G , a set of edges whose removal disconnects the graph is a cut set. If each edge in G has a weight assigned to it, then a cut set with the smallest sum of weights is a minimum cut of the

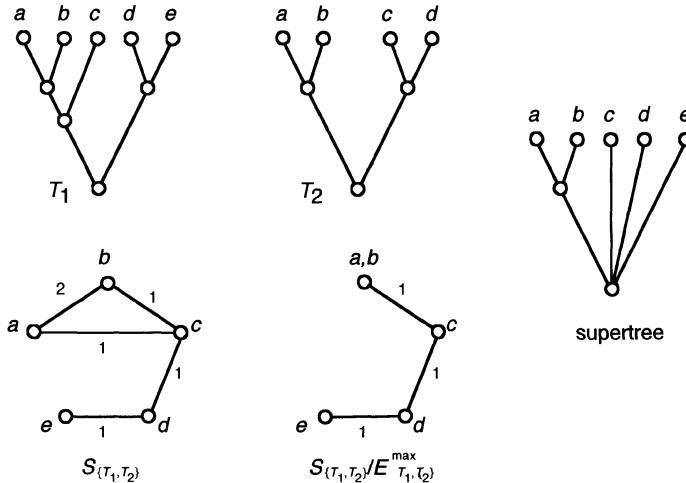


Figure 9. An example of the MINCUTSUPERTREE algorithm showing two input trees T_1 and T_2 , and the graphs S_T / E^{\max}_T . The graph S_T / E^{\max}_T has three minimum cut sets, which yield the components $\{a, b\}$, $\{c\}$, $\{d\}$, and $\{e\}$, which in turn yield a supertree (from Semple and Steel, 2000).

graph. Note that Semple and Steel do not find minimum cuts of S_T , but rather of an associated graph S_T / E^{\max}_T , which they construct as follows:

1. Weight each edge (a, b) in S_T by the number of trees in \mathcal{T} in which a and b are in the same proper cluster.
2. Let E^{\max}_T be the set of edges that have weight k , where k is the number of trees in \mathcal{T} .
3. Merge all nodes in S_T that are connected by edges in E^{\max}_T .

For example, given the two input trees in Figure 9, the edge (a, b) in S_T ensures that all nestings found in all of the input trees \mathcal{T} are in the supertree returned by MINCUTSUPERTREE. It is worth noting that each input tree in \mathcal{T} in Semple and Steel's original algorithm can have a weight $w(T)$ assigned to it, and in step 1 above they weight each edge by the sum of the weights of the trees in which a and b are in the same proper cluster. The set of edges E^{\max} comprises those edges for which $w_{sum} = \sum_{T \in \mathcal{T}} w(T)$. For simplicity here, I consider only the case where all trees have the same unit weight.

4.4.1 Constraints in MINCUTSUPERTREE

The construction of the graph S_T / E^{\max}_T suggests a straightforward method for incorporating constraints. Let T_C be one of the trees in the input set of

trees \mathcal{T} that we designate as the constraint tree. All edges in $S_{\mathcal{T}}$ that arise from T_C are given maximum weight, and consequently will never be in a minimum cut.

4.4.2 Problem: limitations of MINCUTSUPERTREE. Elsewhere (Page, 2002), I constructed an example where MINCUTSUPERTREE gives somewhat paradoxical results, which I attributed to a sensitivity to differences in size among the input trees. We can modify the construction of $S_{\mathcal{T}} / E^{\max}_{\mathcal{T}}$ to reduce its sensitivity to this problem. An implementation of this modification, and Semple and Steel's original algorithm, is available from <http://darwin.zoology.gla.ac.uk/rpage/supertree/>. This modification improves the performance of MINCUTSUPERTREE (results summarized in Burleigh *et al.*, 2004), but it is still less satisfactory than the more computationally demanding MRP and MRF methods. Can the method be modified still further to improve the quality of the supertree it produces without much additional computational effort?

5. Conclusions

Two of the problems, taxonomic names (Section 2.1) and consistency of higher classifications (Section 3.2) are fundamental to assembling meaningfully labeled input trees. The issue of taxonomic overlap (Section 4) is likely to become less of an issue as taxonomic sampling increases. To return to Sanderson's supertree challenge, it is clear that there is some way to go just in terms of data quality before such a challenge can be tackled with a reasonable expectation of getting a meaningful result (never mind the computational issues). This is not to belittle or downplay the computational challenges faced by supertree methods. However, it is a source of some frustration that we cannot assemble large data sets easily on the scale required if we are to assemble the Tree of Life. A major priority for phylogenetics is establishing a phylogenetic database that has properly curated taxonomic names.

Acknowledgements

I would like to thank Olaf Bininda-Emonds for the invitation to contribute to this book, and for waiting patiently for the manuscript to actually arrive. I thank Kevin de Queiroz for his review of the manuscript. Charles Semple also provided helpful comments, and suggested a simpler way to implement constraints than my original approach. Vince Smith listened while I regaled him with complaints about the current state of taxonomy. James Cotton gave

helpful feedback on the manuscript. Christian Lauk assembled the set of bird phylogenies used in this chapter. Some of the ideas presented here were developed while I was on leave in Auckland, New Zealand, in the (northern hemisphere) summer of 2002.

References

- ADAMS, E. M., III. 1986. *N*-trees as nestings: complexity, similarity, and consensus. *Journal of Classification* 3:299–317.
- AHO, A. V., SAGIV, Y., SZYMANSKI, T. G., AND ULLMAN, J. D. 1981. Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. *SIAM Journal of Computing* 10:405–421.
- ALEXE, G., ALEXE, S., FOLDES, S., HAMMER, P. L., AND SIMEONE, B. 2000. *Consensus Algorithms for the Generation of all Maximal Bicliques*. Technical Report 2000–14, DIMACS, Rutgers University, Piscataway, NJ 08854–8018, USA.
- BARRETT, M., DONOGHUE, M. J., AND SOBER, E. 1991. Against consensus. *Systematic Zoology* 40:486–493.
- BAUM, B. R. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon* 41:3–10.
- BAUM, B. R. AND RAGAN, M. A. 2004. The MRP method. In O. R. P. Bininda-Emonds (ed). *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 17–34. Kluwer Academic, Dordrecht, the Netherlands.
- BININDA-EMONDS, O. R. P., JONES, K. E., PRICE, S. A., CARDILLO, M., GRENYER, R., AND PURVIS, A. 2004. Garbage in, garbage out: data issues in supertree construction. In O. R. P. Bininda-Emonds (ed). *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 267–280. Kluwer Academic, Dordrecht, the Netherlands.
- BININDA-EMONDS, O. R. P. AND SANDERSON, M. J. 2001. Assessment of the accuracy of matrix representation with parsimony analysis supertree construction. *Systematic Biology* 50:565–579.
- BRYANT, D. 1997. *Building Trees, Hunting for Trees, and Comparing Trees: Theory and Methods in Phylogenetic Analysis*. Ph.D. dissertation, Department of Mathematics, University of Canterbury.
- BRYANT, D. 2003. A classification of consensus methods for phylogenetics. In M. F. Janowitz, F.-J. Lapointe, F. R. McMorris, B. Mirkin, and F.S. Roberts (eds), *Bioconsensus*, DIMACS: Series in Discrete Mathematics and Theoretical Computer Science, volume 61, pp. 163–183. American Mathematical Society-DIMACS, Providence, RI.
- BURLEIGH, J. G., EULENSTEIN, O., FERNÁNDEZ-BACA, D., AND SANDERSON, M. J. 2004. MRF supertrees. In O. R. P. Bininda-Emonds (ed). *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 65–85. Kluwer Academic, Dordrecht, the Netherlands.
- CHEN, D., EULENSTEIN, O., FERNÁNDEZ-BACA, D., AND SANDERSON, M. J. 2002. *Supertrees by Flipping*. Technical Report TR02–01, Department of Computer Science, Iowa State University, 226 Atanasoff Hall, Ames, IA 50011–1040, USA.
- CHU, P. C. 1995. Phylogenetic reanalysis of Strauch's osteological data set for the charadriiformes. *Condor* 97:174–196.

- CODDINGTON, J. A. 1990. Ontogeny and homology in the male palpus of orb-weaving spiders and their relatives, with comments on phylogeny (Araneoclada: Araneoidea, Deinopoidea). *Smithsonian Contributions to Zoology* 496:1–52.
- CONSTANTINESCU, M. AND SANKOFF, D. 1986. Tree enumeration modulo a consensus. *Journal of Classification* 3:349–56.
- CORMEN, T. H., LEISERSON, C. E., AND RIVEST, R. L. 1990. *Introduction to Algorithms*. The MIT Press, Cambridge, Massachusetts.
- COTTON, J. A. AND PAGE, R. D. M. 2002. Going nuclear: vertebrate phylogeny and gene family evolution reconciled. *Proceedings of the Royal Society of London B* 269: 1555–1561.
- COTTON, J. A. AND PAGE, R. D. M. 2004. Tangled trees from molecular markers: reconciling conflict between phylogenies to build molecular supertrees. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 107–125. Kluwer Academic, Dordrecht, the Netherlands.
- DANIEL, P. AND SEMPLE, C. 2004. A supertree algorithm for nested taxa. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 151–171. Kluwer Academic, Dordrecht, the Netherlands.
- GATESY, J., MATTHEE, C., DESALLE, R., AND HAYASHI, C. 2002. Resolution of a supertree / supermatrix paradox. *Systematic Biology* 51:652–664.
- GATESY, J., O'GRADY, P., AND BAKER, R. H. 1999. Corroboration among data sets in simultaneous analysis: hidden support for phylogenetic relationships among higher level artiodactyl taxa. *Cladistics* 15:271–313.
- GATESY, J. AND SPRINGER, M. S. 2004. A critique of matrix representation with parsimony supertrees. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 369–388. Kluwer Academic, Dordrecht, the Netherlands.
- GOLOBOFF, P. A. AND POL, D. 2002. Semi-strict supertrees. *Cladistics* 18:514–525.
- LAUK, C. 2002. *An Attempt for a Genus-level Supertree of Birds*. B.Sc. (Hons) Project Report, DEEB, IBLS, University of Glasgow.
- LIU, F.-G. R., MIYAMOTO, M. M., FREIRE, N. P., ONG, P. Q., TENNANT, M. R., YOUNG, T. S., AND GUGEL, K. F. 2001. Molecular and morphological supertrees for eutherian (placental) mammals. *Science* 291:1786–1789.
- MIYAMOTO, M. M. 1985. Consensus classifications and general cladograms. *Cladistics* 1:186–189.
- NOVACEK, M. J. 2001. Mammalian phylogeny: genes and supertrees. *Current Biology* 11:R573–R575.
- PAGE, R. D. M. 2002. Modified mincut supertrees. In R. Guigó and D. Gusfield (eds), *Algorithms in Bioinformatics, Second International Workshop, WABI 2002, Rome, Italy, September 17–21, 2002, Proceedings*, pp. 537–552. Springer, Berlin.
- PIEL, W. H., DONOGHUE, M. J., AND SANDERSON, M. J. 2002. TreeBASE: a database of phylogenetic knowledge. In K. Shimura, K. L. Wilson, and D. Gordon (eds), *To the Interoperable Catalogue of Life with Partners — Species 2000 Asia Oceania. Proceedings of 2nd International Workshop of Species 2000*, pp. 41–47. National Institute of Environmental Studies (Research Report R-171–2002), Tsukuba, Japan. (<http://www.nies.go.jp/kanko/kenkyu/pdf/r-171-2002.pdf>)
- POLLOCK, D. D., ZWICKL, D. J., MCGUIRE, J. A., AND HILLIS, D. M. 2002. Increased taxonomic sampling is advantageous for phylogenetic inference. *Systematic Biology* 51:664–671.

- RAGAN, M. A. 1992. Phylogenetic inference based on matrix representation of trees. *Molecular Phylogenetics and Evolution* 1:53–58.
- ROSHAN, U., MORET, B. M. E., WILLIAMS, T. L., AND WARNOW, T. 2004. Performance of supertree methods on various data set decompositions. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 301–328. Kluwer Academic, Dordrecht, the Netherlands.
- ROSS, H. A. AND RODRIGO, A. G. 2004. An assessment of matrix representation with compatibility in supertree construction. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 35–63. Kluwer Academic, Dordrecht, the Netherlands.
- SANDERSON, M. J. 1989. Confidence limits on phylogenies: the bootstrap revisited. *Cladistics* 5:113–129.
- SANDERSON, M. J., PURVIS, A., AND HENZE, C. 1998. Phylogenetic supertrees: assembling the trees of life. *Trends in Ecology and Evolution* 13:105–109.
- SEMPLE, C. AND STEEL, M. 2000. A supertree method for rooted trees. *Discrete Applied Mathematics* 105:147–158.
- SLOWINSKI, J. B. AND PAGE, R. D. M. 1999. How should species phylogenies be inferred from sequence data? *Systematic Biology* 48:814–825.
- STEEL, M. 1992. The complexity of reconstructing trees from qualitative characters and subtrees. *Journal of Classification* 9:91–116.
- SWOFFORD, D. L. 2002. *PAUP**: *Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Version 4. Sinauer, Sunderland, Massachusetts.
- WILKINSON, M., THORLEY, J. L., LITTLEWOOD, D. T. J., AND BRAY, R. A. 2001. Towards a phylogenetic supertree of Platyhelminthes? In D. T. J. Littlewood and R. A. Bray (eds), *Interrelationships of the Platyhelminthes*, pp. 292–301. Taylor and Francis, London.

Chapter 12

GARBAGE IN, GARBAGE OUT

Data issues in supertree construction

Olaf R. P. Bininda-Emonds, Kate E. Jones, Samantha A. Price, Marcel Cardillo, Richard Grenyer, and Andy Purvis

Abstract: As in conventional phylogenetic analyses, issues surrounding the source data are paramount in the supertree construction, but have received insufficient attention. In supertree construction, however, the source data represent phylogenetic trees rather than primary character data. This presents several supertree-specific problems. In this paper, we examine several key data issues for supertree construction, including data set non-independence, taxonomy of terminal taxa, and the question of what constitutes a valid source tree. Throughout, we present our suggested protocol for source tree collection and manipulation based on our experiences in building a supertree of mammals. Other protocols and decisions are naturally possible. What is important is that all collection protocols are presented explicitly and address minimally the issues that we have identified.

Keywords: character data; data non-independence; monophyly; paraphyly; source trees

1. Introduction

Supertree construction represents a class of techniques in which at least partially overlapping evolutionary trees are combined to produce a single (usually) more comprehensive tree, the supertree. It differs from most other forms of phylogenetic analysis in that the raw data comprise tree topologies rather than traditional morphological or molecular characters. Despite this important distinction, most of the theoretical research into supertrees has focused on assessing the performance of existing methods (both empirically and from first principles) and on developing new methods (for a review, see Bininda-Emonds *et al.*, 2002). Apart from a few papers (e.g., Springer and

de Jong, 2001; Gatesy *et al.*, 2002; Gatesy and Springer, 2004), little attention has been paid to the raw data of a supertree analysis, namely the source trees that are combined into the supertree. This state of affairs parallels the situation in traditional, character-based phylogenetics, where discussions of the issues involving character selection and definition are comparatively rare (however, see Jenner, 2001).

Issues concerning source trees are of fundamental importance to supertree construction and transcend the different supertree methods (as well as applying to consensus techniques). Different supertree methods possess different properties (see Wilkinson *et al.*, 2004) and have better or worse performance under certain circumstances, but they are all limited ultimately by characteristics of the source trees that are being combined. An especially important issue is that of source tree non-independence, whereby the same primary character data contributes to more than one source tree (Springer and de Jong, 2001; Gatesy *et al.*, 2002).

In this paper, we examine issues related to the collection and manipulation of source trees as part of a supertree analysis. We also provide a suggested protocol based on our experiences in building a supertree of all extant species of mammal. Parts of this protocol will apply to all supertree methods, whereas others will be specific to matrix-based supertree methods such as matrix representation with parsimony (MRP; Baum, 1992; Ragan, 1992). Although the protocol does not resolve all issues involved in building supertrees, we believe it to be the best working protocol and one that is based on explicit procedures such that it can be applied easily by different researchers to generate the same end results.

2. Issues concerning source trees

2.1 Non-independence and duplication

It is common practice in phylogenetic systematics for characters (and often the character states) to be obtained from the literature and re-used in a novel analysis. This is true for both morphological (see Jenner, 2001) and molecular data. For example, in the early days of DNA-sequence analysis, individual research groups often published a series of papers, each of which in turn included newly sequenced species that were added to the base data set. As a result, the same basic piece of character information can contribute to more than one source tree, resulting in data duplication in a supertree analysis. In all cases of data duplication, the overlap of character data between source studies means that the associated source trees are not independent of one another, a key assumption of phylogenetic analysis.

Moreover, the degree of non-independence and duplication is often difficult to quantify between any two source trees and virtually impossible for even a modest set of interrelated source trees (but see Gatesy *et al.*, 2002).

Although most attention has focused on non-independence among trees from different papers (Springer and de Jong, 2001; Gatesy *et al.*, 2002; Gatesy and Springer, 2004), non-independence can also arise among trees presented within any single source study (between- and within-study non-independence, respectively). The latter form of non-independence arises because individual data sets are often analyzed using multiple optimization criteria and weighting schemes; using different subsets of the data (both characters and taxa); and, for molecular sequence data, using different alignments. Thus, a given study can present numerous potential source trees that are clearly not independent of one another.

Unless it is accounted for, data non-independence and the associated data duplication means that some data sets are effectively upweighted and might have more influence on the supertree analysis. For instance, both Springer and de Jong (2001) and Gatesy *et al.* (2002) point out examples of single data sources being replicated many times in the supertree analysis of Liu *et al.* (2001). Similar instances of data duplication occur undoubtedly in most of the published supertrees (exceptions include Daubin *et al.*, 2001, 2002; Kennedy and Page, 2002), despite steps taken to minimize them.

Data set non-independence is arguably the greatest problem facing supertree construction and all methods of combining trees. This problem usually cannot be eliminated entirely, and in fact will become more of an issue with the increasing number of total evidence studies and the re-use of data, particularly molecular data, facilitated by on-line archiving of data sets (e.g., GenBank, web pages of individual journals, or TreeBASE; Sanderson *et al.*, 1994; Piel *et al.*, 2002). However, we feel that data set non-independence can be largely ameliorated using an appropriate source tree collection protocol. This was shown indirectly by Gatesy *et al.* (2002) in their re-analysis of the Liu *et al.* (2001) mammal supertree analysis. When Gatesy *et al.* pruned out source trees that they felt were redundant or poorly justified, they recovered a supertree containing Cetacea within a paraphyletic Artiodactyla (i.e., they recovered the clade Hippopotamidae + Cetacea), a result that they felt was more in accord with current systematic opinion. However, the reworked matrix was still held to contain redundant information (J. Gatesy, pers. comm.).

2.1.1 Identifying independent source trees

The guiding rule in our protocol, which is summarized in Figure 1, is to identify phylogenetic hypotheses that can be viewed reasonably as

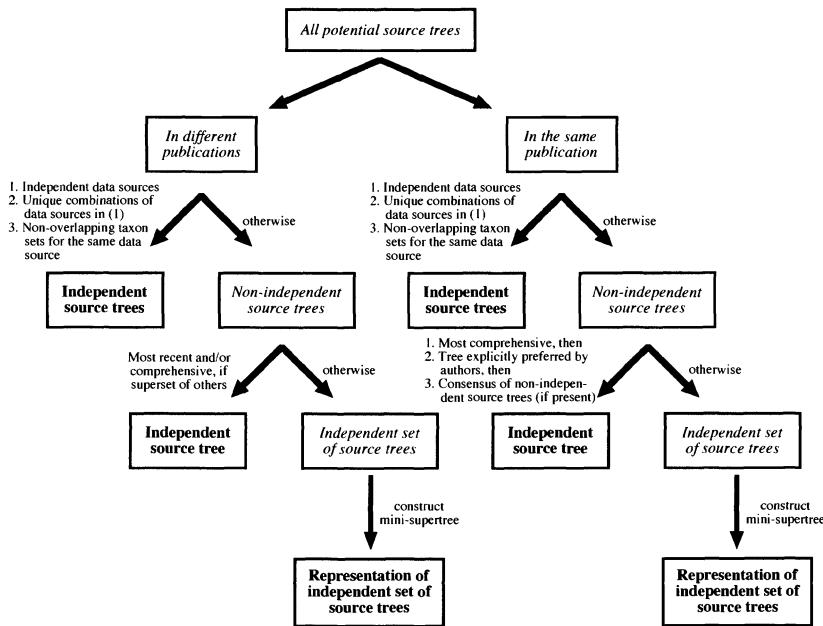


Figure 1. Decision tree summarizing our suggested protocol for source tree collection. Boxed entries in italics require further processing. Boxed entries in bold face represent reasonably independent units that can be included in a supertree analysis. Source trees can be rejected at any level (not shown).

independent (following Purvis, 1995b; Bininda-Emonds *et al.*, 2003). Independence is a difficult concept about which to be precise in phylogenetics, where the aim is to recover the common hierarchical set of relationships that has underpinned the evolution of all characters. Hence, characters are not independent in the sense of being generated by independent processes. Given the acceptance of gene trees (versus species trees; Maddison, 1997), the recognition of pseudo-independent evolutionary “packets” (i.e., the genes, possibly down to their individual exons; and the trees derived from them) might be defensible, and accords with the original justification for MRP as a method to combine gene trees (Baum, and Ragan, 2004). Our protocol attempts to identify these packets and is based on an explicitly defined set of rules for deciding the precedence of one tree over another. These rules have been formulated according to the criteria of 1) data independence, 2) taxonomic inclusiveness, and 3) (informed) author preference.

We base decisions about independence on both the source of the character data and the taxon set, not on the publications in which they appear. Non-overlapping data sets (e.g., different genes) are considered to be

independent data sources, even if they appear on a single heritable unit like mitochondrial DNA. However, different portions of the same gene (and possibly each exon within a gene), because of their common evolutionary history, are not independent for an overlapping set of taxa, even if these gene portions do not overlap at all. Trees for non-overlapping taxon sets, even if they are derived from the same set of characters, are independent by practical necessity: there is no way to combine such data sets meaningfully other than to combine the primary character data.

We also hold unique combinations of genes to be independent sources and independent from data sets containing subsets of all the genes in the combination (i.e., a total evidence analysis; *sensu* Kluge, 1989). We base this decision on the phenomenon of “signal enhancement” (*sensu* de Queiroz *et al.*, 1995), whereby the combination of data sets can yield a novel solution that is not indicated by any of the constituent data sets (see Barrett *et al.*, 1991). In essence, we hold that signal enhancement causes total evidence solutions to constitute independent phylogenetic hypotheses, although they might be based on data used elsewhere. We would argue that different morphological data sets are equivalent to novel combinations of genes and so are considered to be independent of one another unless one data set is contained completely within another.

As mentioned, non-independence can exist between or within studies. When it is present between studies, we suggest using only the most recent and/or most comprehensive study (in terms of number of taxa), but only if this study is a superset of all other source trees. This is true whether the same or different research groups have published the studies. Where no single choice presents itself (e.g., no study exists that contains all the taxa found in the others), all equally suitable source trees should be collected for a subsequent, intermediate analysis (see below).

For within-study non-independence, the first step is to identify independent sets of trees based on data set independence. For each such set, the first choice of source tree is the most comprehensive one, based on both taxa and characters. Where this is not possible, the next choice is the phylogenetic hypothesis that is preferred explicitly by the authors of the study. If no single tree is preferred clearly, the (preferred) consensus of all the trees in the set should be used instead. Finally, should multiple equally suitable source trees remain, all should be collected.

2.1.2 Accommodating non-independent sets of source trees

Often it will be possible to select a single source tree to represent a set of non-independent source trees (e.g., the most comprehensive source tree or the one preferred explicitly by the author(s) of a study). When this is not

possible, the set of non-independent source trees should be treated as a single unit. However, these units can still possess a disproportionate influence on the analysis because they are not single trees, but sets of trees. Three solutions present themselves to correct for this. The first two apply only to matrix representation supertree methods.

First, the source trees in each unit can be downweighted such that the unit as a whole has the same weight as a single source tree. However, this solution will not be feasible usually because it is not clear in many cases what the corrected weight should be given that individual source trees will themselves have different weights (i.e., numbers of matrix elements) according to their size and resolution. Second, as suggested by Bininda-Emonds and Bryant (1998) for coding in a set of equally most parsimonious solutions, only each unique node in the set of source trees should be coded. By itself, this yields a solution identical with the strict consensus of the set of source trees. However, it also retains the conflicting clustering information that is otherwise subsumed in the strict consensus tree. This solution is possible only if unique nodes can be identified unambiguously. Thus, the set of source trees must have identical taxon sets and information regarding node support cannot be included. Third, one can produce a “mini-supertree” for the set of source trees, and then use this mini-supertree as the source tree in the main analysis (e.g., Bininda-Emonds *et al.*, 1999). This procedure can be used with all supertree algorithms and on sets of source trees that do not have identical taxon sets. However, there is a loss of information in that the mini-supertree usually will not display all the relationships present in the set of source trees. Moreover, the mini-supertree will be produced often under conditions where MRP methods at least have been demonstrated to possess reduced power (i.e., few source trees; Bininda-Emonds and Sanderson, 2001). Despite this, we recommend the use of mini-supertrees over the other two options because of its generality and ease of application.

2.2 Standardizing terminal taxa

In all studies where data are combined, be they primary character data or source trees, one needs to ensure that the terminal taxa, be they fossil or extant, are comparable throughout the source data (also Page, 2004). This is essential for terminal taxa that appear in more than one source data set. However, taxonomic differences between the data sets will often make the required assessments difficult. A useful solution, therefore, is to standardize the taxonomy of the terminal taxa. In so doing, all taxonomic information from the source study should be collected and retained. This will simplify the standardization process as well as facilitate the possible use of different taxonomic systems in future analyses.

Many of the ideas in this section have been implemented in the Perl script synonoTree.pl, available from <http://www.tierzucht.tum.de/Bininda-Emonds> or from the first author.

2.2.1 Combining trees at different taxonomic levels

The terminal taxa of different source trees can be species (or subspecies) or from higher taxonomic levels. The latter situation is more common in (older) morphological studies, where the states described usually do not refer to an actual species, but to the inferred ancestral (or groundplan) states of the higher taxon (e.g., Wyss and Flynn, 1993). The use of higher taxa in molecular studies is less common and derives usually from a given species being held to be an exemplar for (i.e., representative of all members of) a higher-level group.

It is possible to build supertrees with the terminal taxa representing different taxonomic levels. For example, the top-level supertree of the Carnivora (figure 1 of Bininda-Emonds *et al.*, 1999) contains both species (the red panda, *Ailurus fulgens* and the walrus, *Odobenus rosmarus*) and numerous families. Similarly, the Liu *et al.* (2001) mammal supertree contains both orders (Carnivora and Primates) and families (remaining terminal taxa). This is valid in both cases because the terminal taxa are not nested hierarchically and so are independent of one another.

Where the terminal taxa are nested, the use of classification graphs (see Page, 2004) can allow the source trees to be combined, but only if they have labeled internal nodes (see also Daniel and Semple, 2004). A more universal solution is to standardize the names using either the higher- or lower-level names. The decision on which level to use depends obviously on the desired taxonomic level of the supertree. In both cases, the required decisions make important taxonomic assumptions and thus should be made according to a standard, well-recognized taxonomic reference containing sufficient information upon which to base them (e.g., synonymy lists and type localities).

If the higher-level name is to be used, then all constituent lower-level taxa should take on this name (i.e., essentially taken to be exemplars of the higher taxon), with all monophyletic clades being collapsed to a single terminal. This procedure can occasionally cause the higher taxon to be non-monophyletic in some source trees. Our solution to this problem appears in the following section. This procedure also makes a strong assumption regarding the monophyly of the higher-level taxon, which can cause erroneous results if it is wrong (Gatesy *et al.*, 2002; Malia *et al.*, 2003). It is also important to ensure that the higher-level taxon is comparable (i.e., that a consistent definition is used) across the source studies.

When using the lower-level name, the first step is to identify the actual species of the higher taxon that were examined in the source study. This is the most desirable, direct, and least assumption-laden option. Two solutions present themselves if this is not possible. First, one can assume the monophyly of the higher taxon and create an extra node consisting of all the species (or relevant taxonomic units) in the higher taxon. However, this procedure will elevate the support for the monophyly of the higher taxon artificially according to the membership of the taxonomic reference. As such, the support derives from an appeal to authority (*sensu* Gatesy *et al.*, 2002) and not hard evidence. Moreover, the current content of the higher-level group might differ from that recognized at the time of the source study as a result of changes in phylogenetic opinion or simply the use of a different taxonomic system. Instead, we suggest that the type species of the higher taxon be identified and used in its place (following Jones *et al.*, 2002). Although this procedure is also assumption-laden, it makes fewer assumptions of monophyly and thus influences the topology of the supertree to a lesser degree.

2.2.2 Non-monophyletic species

The ever-changing and contentious nature of taxonomy and species assignment means that source studies are often not comparable in the species that they recognize (i.e., they use different synonyms or even synonyms that are no longer valid). It is therefore desirable to standardize the species taxonomy using a single, recognized reference. However, doing so can result in the creation of non-monophyletic species (Figure 2). For example, a given source study might recognize X and Y as being distinct species, and ones that are not each other's closest relatives. By contrast, the reference taxonomy recognizes X and Y as being the same species, Z, rendering Z non-monophyletic in that source study. For clarity, we refer to X and Y as the “source species” and Z as the “reference species” in the following.

Two solutions present themselves. First, as for coding higher-level taxa, the source species that represents the type species in the reference taxonomy should be recognized as the reference species. Second, if this is not possible (e.g., the type species is unknown or neither source species can be equated with it), then the position of the reference species should be considered as being uncertain in that source tree. In essence, the single source tree represents multiple source trees, one for each source species (now representing the reference species) with the remaining source species being pruned (Figure 2b). These multiple non-independent source trees can be handled in the normal fashion. Either the source trees can be used to form a

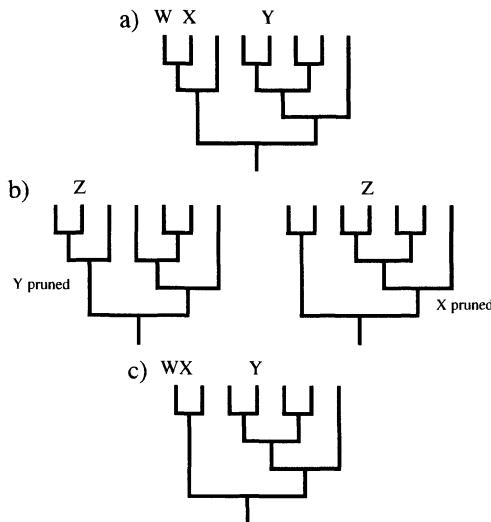


Figure 2. The problem of non-monophyletic species. The reference taxonomy holds the source species X and Y to be the same species, Z, rendering Z polyphyletic on the tree in (a). One possible solution is to prune each of the source species in turn to yield a set of source trees reflecting the uncertain placement of Z (b). If two or more source species of Z form a monophyletic clade (e.g., W and X in (a)), this clade can be collapsed to a single terminal (c).

mini-supertree or, for matrix representation methods, each unique node among the set of source trees can be coded.

It is also possible that a source study will have a reference species formed from a combination of monophyletic and non-monophyletic source species. For example, in Figure 2a, consider that the sister species of X, W, is also a source species of Z. In this case, W and X can be collapsed into the single source species WX (Figure 2c). However, both W and X should be considered separately in WX when attempting to determine if either is the type species of Z.

2.3 Valid versus invalid source trees

A multitude of phylogenies derived from a variety of data sources and methodologies exist in the systematic literature. Older source trees in particular are often not based on any explicit data source or methodology. Obviously, all these potential source trees can differ widely in quality and potential utility.

The decision as to what constitutes a valid (or usable) source tree has differed in past supertree analyses. These decisions have been guided by the

often-conflicting issues of data quality, achieving sufficient taxonomic coverage, and the goal of the supertree analysis. For example, Pisani *et al.* (2002) explicitly collected source trees of the Dinosauria that were obtained using cladistic methods, and therefore postdate 1966. By contrast, Purvis (1995a) included all phylogenetic hypotheses available, even those that were never intended as such, to obtain complete taxonomic coverage of the Primates. However, he acknowledged the difference in source tree quality by downweighting those trees that were not derived using a robust methodology heavily. Finally, in trying to summarize the prevalent systematic opinion for a given historical period, Bininda-Emonds (in press) included all available source trees and weighted them equally.

When the goal of the supertree study is to derive the best possible estimate of the phylogeny of a given group, only source trees of the highest available quality should be used (Gatesy *et al.*, 2002; Bininda-Emonds *et al.*, 2003), within the constraints of being able to assess this quality *a priori* and achieving the desired taxonomic coverage. The deleterious effects of including poor quality data manifest themselves probably only if wrong statements from these trees act in concert and, more importantly, if there are few good data. Thus, data quality control is particularly important when there are relatively few source trees. With many source trees, the inclusion of lesser quality data should have little effect on the result. However, one should be mindful that there is likely to be a trade-off between source tree number and quality. Implicit in this discussion is that poor quality data are misleading. In one case study, however, poor and good quality data produced estimates of phylogeny that were indistinguishable from one another statistically (Bininda-Emonds, 2000). This finding obviously need not be representative of all taxonomic groups or analyses, however.

Our suggestion is that only source trees based on original analyses should be collected. Secondary representations of a source tree from another study should be deferred to the tree in the original study. It is debatable whether or not taxonomies or even other supertrees constitute original analyses. Both represent secondary manipulations of original data, but also potentially novel phylogenetic hypotheses. In our work, we have included the former (unless the phylogeny that it is based on is clear), but not the latter. Instead, we collected the source trees from which the supertree was derived.

On a more practical note, we also recommend that only trees from published sources, or minimally ones that are “in press”, be collected. Doing so increases the accountability (*sensu* Gatesy *et al.*, 2002) of the supertree analysis with respect to its source data. Unpublished data are liable to change before they are published, might be published in another format or venue and therefore be untraceable, or might never be published. We would include web journals and, so long as the results have not been published elsewhere,

graduate dissertations as valid published sources. Although the latter are not peer reviewed in the strict sense, the key issues here are data accessibility and accountability. The inferred quality of the source trees is an important, but separate consideration.

In the end, the researcher must make the final decision as to what represents a valid source tree. However, it should be made according to explicit guidelines, based on, but not limited to, issues such as data quality, the methodology used, and assumptions made in the study (e.g., appeals to authority or other assumptions of monophyly). Moreover, we urge the use of differential weighting to explore the effect of source tree quality on the supertree analysis (e.g., Purvis, 1995a; Bininda-Emonds *et al.*, 1999; Jones *et al.*, 2002; Stoner *et al.*, 2003).

2.4 Source-tree collection

In all cases, we recommend that suitable source trees be collected exactly as they appear in the source study, together with all important taxonomic information for the taxa that appear on them. Copies of the source trees can then be modified to suit the particular supertree analysis, providing maximum flexibility. For example, the source trees can be pruned to include only the set of taxa of interest (e.g., pruning fossil species) or the taxon names can be standardized to accord with a given taxonomic reference or the desired taxonomic level of the supertree.

3. Conclusion

The protocol that we outline above represents a refinement and formalization of the different procedures used previously in supertree construction (Purvis, 1995a; Bininda-Emonds *et al.*, 1999; Liu *et al.*, 2001; Jones *et al.*, 2002; Stoner *et al.*, 2003) and has been useful in our efforts to build a complete species-level mammal supertree. The protocol should not be taken to be the absolute solution to supertree construction, but instead be interpreted as some guidelines to a set of defined problems. Different solutions can, and probably will, be necessary for supertrees of other groups and depending on the goal of the supertree study. What is important for every supertree analysis is that the rules used to select and manipulate the source trees are made explicit to allow reconstruction of the results.

Acknowledgements

We thank Harold Bryant and John Gatesy for their helpful comments. This work was conducted as part of the “Phylogeny and Conservation” Working Group supported by the National Center for Ecological Analysis and Synthesis, a center funded by NSF (grant DEB-94-21535), the University of California at Santa Barbara, and the State of California. Additional support was through the German research program BMBF (OBE), a NERC studentship GT04 1999/TS/0140 (RG), NERC grant NER/A/S/2001/000581 (MC and AP) and NSF grant DEB/0129009 (KEJ and SAP). This work was completed while KEJ was at the Department of Biology, University of Virginia and RG was at the Department of Biological Sciences, Imperial College.

References

- BARRETT, M., DONOGHUE, M. J., AND SOBER, E. 1991. Against consensus. *Systematic Zoology* 40:486–493.
- BAUM, B. R. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon* 41:3–10.
- BAUM, B. R. AND RAGAN, M. A. 2004. The MRP method. In O. R. P. Bininda-Emonds (ed). *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 17–34. Kluwer Academic, Dordrecht, the Netherlands.
- BININDA-EMONDS, O. R. P. 2000. Factors influencing phylogenetic inference: a case study using the mammalian carnivores. *Molecular Phylogenetics and Evolution* 16:113–126.
- BININDA-EMONDS, O. R. P. In press. The phylogenetic position of the giant panda (*Ailuropoda melanoleuca*): a historical consensus through supertree analysis. In D. G. Lindburg and K. Baragona (eds), *Pandas: Biology and Conservation*. University of California Press, Berkeley.
- BININDA-EMONDS, O. R. P. AND BRYANT, H. N. 1998. Properties of matrix representation with parsimony analyses. *Systematic Biology* 47:497–508.
- BININDA-EMONDS, O. R. P., GITTLEMAN, J. L., AND PURVIS, A. 1999. Building large trees by combining phylogenetic information: a complete phylogeny of the extant Carnivora (Mammalia). *Biological Reviews* 74:143–175.
- BININDA-EMONDS, O. R. P., GITTLEMAN, J. L., AND STEEL, M. A. 2002. The (super)tree of life: procedures, problems, and prospects. *Annual Review of Ecology and Systematics* 33:265–289.
- BININDA-EMONDS, O. R. P., JONES, K. E., PRICE, S. A., GRENYER, R., CARDILLO, M., HABIB, M., PURVIS, A., AND GITTLEMAN, J. L. 2003. Supertrees are a necessary not-so-evil: a comment on Gatesy *et al.* *Systematic Biology* 52:724–729.
- BININDA-EMONDS, O. R. P. AND SANDERSON, M. J. 2001. Assessment of the accuracy of matrix representation with parsimony supertree construction. *Systematic Biology* 50:565–579.
- DANIEL, P. AND SEMPLE, C. 2004. A supertree algorithm for nested taxa. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 151–171. Kluwer Academic, Dordrecht, the Netherlands.

- DAUBIN, V., GOUY, M., AND PERRIÈRE, G. 2001. Bacterial molecular phylogeny using supertree approach. *Genome Informatics* 12:155–164.
- DAUBIN, V., GOUY, M., AND PERRIÈRE, G. 2002. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Research* 12:1080–1090.
- DE QUEIROZ, A., DONOGHUE, M. J., AND KIM, J. 1995. Separate versus combined analysis of phylogenetic evidence. *Annual Review of Ecology and Systematics* 26:657–681.
- GATESY, J., MATTHEE, C., DESALLE, R., AND HAYASHI, C. 2002. Resolution of a supertree / supermatrix paradox. *Systematic Biology* 51:652–664.
- GATESY, J. AND SPRINGER, M. S. 2004. A critique of matrix representation with parsimony supertrees. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 369–388. Kluwer Academic, Dordrecht, the Netherlands.
- JENNER, R. A. 2001. Bilaterian phylogeny and uncritical recycling of morphological data sets. *Systematic Biology* 50:730–742.
- JONES, K. E., PURVIS, A., MACLARNON, A., BININDA-EMMONDS, O. R. P., AND SIMMONS, N. B. 2002. A phylogenetic supertree of the bats (Mammalia: Chiroptera). *Biological Reviews* 77:223–259.
- KENNEDY, M. AND PAGE, R. D. M. 2002. Seabird supertrees: combining partial estimates of procellariiform phylogeny. *The Auk* 119:88–108.
- KLUGE, A. G. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among Epicrates (Boidae, Serpentes). *Systematic Zoology* 38:7–25.
- LIU, F.-G. R., MIYAMOTO, M. M., FREIRE, N. P., ONG, P. Q., TENNANT, M. R., YOUNG, T. S., AND GUGEL, K. F. 2001. Molecular and morphological supertrees for eutherian (placental) mammals. *Science* 291:1786–1789.
- MADDISON, W. P. 1997. Gene trees in species trees. *Systematic Biology* 46:523–536.
- MALIA, M. J., JR., LIPSCOMB, D. L., AND ALLARD, M. W. 2003. The misleading effects of composite taxa in supermatrices. *Molecular Phylogenetics and Evolution* 27:522–527.
- PAGE, R. D. M. 2004. Taxonomy, supertrees, and the Tree of Life. In O. R. P. Bininda-Emonds (ed.). *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 247–265. Kluwer Academic, Dordrecht, the Netherlands.
- PIEL, W. H., DONOGHUE, M. J., AND SANDERSON, M. J. 2002. TreeBASE: a database of phylogenetic knowledge. In K. Shimura, K. L. Wilson, and D. Gordon (eds), *To the Interoperable Catalogue of Life with Partners — Species 2000 Asia Oceania. Proceedings of 2nd International Workshop of Species 2000*, pp. 41–47. National Institute of Environmental Studies (Research Report R-171-2002), Tsukuba, Japan. (<http://www.nies.go.jp/kanko/kenkyu/pdf/r-171-2002.pdf>)
- PISANI, D., YATES, A. M., LANGER, M. C., AND BENTON, M. J. 2002. A genus-level supertree of the Dinosauria. *Proceedings of the Royal Society of London B* 269:915–921.
- PURVIS, A. 1995a. A composite estimate of primate phylogeny. *Philosophical Transactions of the Royal Society of London B* 348:405–421.
- PURVIS, A. 1995b. A modification to Baum and Ragan's method for combining phylogenetic trees. *Systematic Biology* 44:251–255.
- RAGAN, M. A. 1992. Phylogenetic inference based on matrix representation of trees. *Molecular Phylogenetics and Evolution* 1:53–58.
- SANDERSON, M. J., DONOGHUE, M. J., PIEL, W., AND ERIKSSON, T. 1994. TreeBASE: a prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life. *American Journal of Botany* 81:183.

- SPRINGER, M. S. AND DE JONG, W. W. 2001. Phylogenetics. Which mammalian supertree to bark up? *Science* 291:1709–1711.
- STONER, C. J., BININDA-EMONDS, O. R. P., AND CARO, T. M. 2003. The adaptive significance of coloration in lagomorphs. *Biological Journal of the Linnean Society* 79:309–328.
- WILKINSON, M., THORLEY, J. L., PISANI, D., LAPOINTE, F.J., AND MCINERNEY, J. O. 2004. Some desiderata for liberal supertrees. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 227–246. Kluwer Academic, Dordrecht, the Netherlands.
- WYSS, A. R. AND FLYNN, J. J. 1993. A phylogenetic analysis and definition of the Carnivora. In F. S. Szalay, M. J. Novacek, and M. C. McKenna (eds), *Mammalian Phylogeny: Placentals*, pp. 32–52. Springer-Verlag, New York.

Chapter 13

RECONSTRUCTING DIVERGENCE TIMES FOR SUPERTREES

A molecular approach

Rutger A. Vos and Arne Ø. Mooers

Abstract: Here, we present a formal approach to estimating divergence dates derived from aligned DNA sequence data on MRP supertrees, using a new supertree for the Primates as a case study. We selected 40 sequence data sets that conform under various models of sequence evolution to the molecular clock. Each of these data sets covers only a subset of the taxa on the supertree, and so composite date estimates were obtained by calibrating the data sets on common nodes and subsequently combining the estimates from different genes for the same node. The internal consistency of our estimates is high. The estimates presented here also fit well with those from Purvis' 1995 primate supertree, although estimates for deeper splits are progressively older.

Keywords: divergence times; fossils; maximum likelihood; molecular clock; Primates; supertree techniques

1. Introduction

Supertrees can be applied usefully to research beyond that of descriptive systematics (Bininda-Emonds *et al.*, 2002; Gittleman *et al.*, 2004), including comparative studies of character evolution (Gittleman *et al.*, 2004); studies of speciation, extinction, and diversification rates (Purvis *et al.*, 1995; Moore *et al.*, 2004); or establishing conservation priorities (e.g., based on the “evolutionary heritage” concept, the amount of independent evolutionary history embodied within a taxon; Mooers *et al.*, in press). These applications require phylogenies for which divergence dates, relative or absolute, are established. Although estimates of relative branch lengths from consensus

techniques are possible (see Bryant *et al.*, 2004), the most widely used technique for the amalgamation of source trees, matrix representation with parsimony analysis (MRP; Baum, 1992; Ragan 1992), does not result in branch lengths that can be interpreted as a temporal dimension. Instead, divergence dates on such supertrees are added afterwards, if at all. In some of the currently published supertrees, divergence dates were obtained through a combination of fossil dates, indirect estimates of sequence divergence by measuring branch lengths from published sources, and models for the expected age of clades given the number of taxa of that clade relative to its dated parent clade (Purvis, 1995; Bininda-Emonds *et al.*, 1999). In other studies (e.g., Wojciechowski *et al.*, 2000; Liu *et al.*, 2001; Jones *et al.*, 2002; Pisani *et al.*, 2002), no effort was made to establish divergence dates. In any case, objective and robust methods to reconstruct divergence dates for MRP supertrees directly from molecular data sets have yet to be established. Here, we will comment on the advantages and pitfalls of different techniques and data sources, and then discuss a molecular approach as applied to a new supertree of the order Primates.

1.1 Fossils as tools for calibration

If a fossil can be ascribed clearly to a clade, it can offer a minimum estimate of the age of that clade. The application of fossils in estimating divergence dates is twofold: a fossil date can not only be used to define the minimal age of a single node or clade (and its sister group) in a tree, but also to calibrate the absolute depths of other nodes in the same tree if the relative depths of these nodes have been inferred (e.g., from gene sequence data). This distinction is worth mentioning in the context of supertrees: relative node ages are unknown for MRP supertrees and fossils can supply information only in the former manner (i.e., as an indicator of the minimum age of clades and their sister groups) without recourse to added data. The paucity of fossil data is therefore an especially big problem in this type of supertree construction and subsequent dating.

The data on the ages of taxa provided by the fossil record has conflicted with molecular phylogenetic data on several occasions. A textbook example of such conflict is the initial identification of *Ramapithecus* as a 9–12 million year old hominid, constraining the split between humans and the (non-hominid) chimpanzees to be older than that. The subsequent reclassification of *Ramapithecus* as being more closely related to orangutans reconciled the fossil-constrained age of the hominids with the mounting molecular evidence of a more recent origin (Ridley, 1996). Clearly, a misidentified fossil leads to correlated errors for all the node depth calibrations based on it. The reliability and independence of fossil dates

should therefore be evaluated critically, as stressed by, for example Lee (1999), who showed that recent molecular evidence for the earliest metazoan split (Xun, 1998) was calibrated on only two “fossil” dates — one of which was actually obtained from the other “with an additional (molecular) layer of uncertainty introduced” (Lee, 1999:387).

The carnivore supertree (Bininda-Emonds *et al.*, 1999) is an example where fossils were used to derive the minimum age of sister groups: the time of first occurrence of either descendant lineage was used to date nodes. It is agreed generally that because fossils can be classified only once clade-defining morphological synapomorphies have arisen (Archibald, 1999), it is likely that the fossils of the earliest members of a clade are often overlooked as members of the clade (if these fossils have formed and were discovered at all). Thus, fossil dates will be too-young estimates of the age of clades. A famous example of this is the “Cambrian explosion” scenario, a hypothesized evolutionary burst (e.g., Gould, 1989; Lipps and Signor, 1992) that hinges on the assumption that sudden cladogenesis and trait evolution followed from the sudden appearance of most animal phyla in the Cambrian fossil record. Molecular studies, however, consistently support an extended period of Precambrian metazoan diversification (Bromham *et al.*, 1998; Bromham and Hendy 2000) along “ghost lineages” (Novacek and Wheeler, 1992; Fortey *et al.*, 1996), giving further evidence that fossils should not be considered as fixed ages of nodes, but rather as constraints on the minimum ages of nodes. However, despite the difficulties in working with fossils in terms of their rarity and their interpretation, the key attraction to fossils is that they are the only way, ultimately, that absolute ages of clades can be determined.

1.2 Relative divergence dates inferred from molecular phylogenies

DNA sequence data can provide information on when species have diverged, not only on the branching order that can be inferred from the phylogenetic signal they provide, but also on the relative timing of these branching events. For the latter to work, the locus under study must conform to the “molecular clock” (Zuckercandl and Pauling, 1965), which in practice means that substitution rates must be constant along all lineages, resulting in an ultrametric tree (i.e., a tree with the same root-to-tip path length for all lineages). Whether or not a particular locus conforms to the molecular clock can be tested by comparing the likelihood (Felsenstein, 1981) of the optimal topology under unconstrained rates to the likelihood of the same tree constrained to be ultrametric. The ultrametric tree will have a worse score,

but, if it is not significantly worse, the locus is considered to conform to the molecular clock hypothesis.

Clocklike loci are a useful source of information from which divergence dates for supertrees can be obtained. However, the MRP supertree technique does not allow for branch length information to be encoded in such a way that the resulting supertree reproduces meaningful divergence date estimates. Therefore, in earlier MRP supertree studies, molecular data on divergence times was used indirectly (Purvis, 1995; Bininda-Emonds *et al.*, 1999) by rescaling previously published molecular phylogenies, calibrating them subsequently using fossil data, and then sticking the divergence dates so obtained on the supertree. This approach has two drawbacks. First, the rescaling process (as described in Purvis, 1995) is essentially a method by which source phylogenies are “ultrametricized” without recourse to the underlying sequence data. It is therefore not certain whether or not the particular locus actually conforms to the molecular clock. Second, the source trees sometimes do not match the topology of the supertree, rendering the source tree in whole or in part unusable. Given these drawbacks, we argue that using sequence data directly is an approach that warrants further research, a case study of which is discussed in this chapter.

1.3 Obtaining composite estimates of divergence dates from sequence data

Relative branch lengths from a set of congruent phylogenies that each cover a subset of taxa usually cannot be combined to derive the branch lengths for the phylogeny that covers the bigger set from which the subsets were drawn. However, the depths of nodes in the set of congruent phylogenies can be combined. For instance, Figure 1a shows a topology for which the divergence dates are unknown. The four trees in Figure 1b each cover a subset of the taxa of the tree in Figure 1a, and are congruent with the topology of that tree. The branch lengths for these ultrametric trees might have been derived from disparate data sources, such as different genes that conform to the molecular clock hypothesis. By calibrating these trees on a shared node — such as node 2 for trees II–IV in this example — the node depths of these trees can be combined to obtain the branch lengths for the topology of Figure 1a (as shown to the right in Figure 1c). From the example in Figure 1, it is evident that this method can be used only to combine divergent dates from multiple sources that share at least one node. However, this is not the only consideration that needs to be taken into account in choosing calibration points.

The location of the calibration point relative to the other nodes in the source trees has an effect on how variation in the estimates is distributed

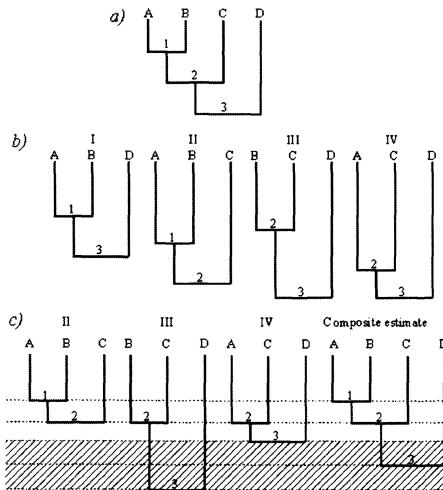


Figure 1. Combining and calibrating divergence dates. a) Hypothetical MRP supertree topology, for which relative branch lengths and labeled node depths are undefined. b) Aligned sequence data sets that conform to the molecular clock when fitted to the topology of the supertree. Node labels correspond with those in (a). c) Sequence data sets II, III and IV are calibrated on their shared node 2. Based on these combined data sets, the depths for all three nodes can be reconstructed in the composite estimate. Because there are two data points for node 3, there is a range (hatched area) from which the median is selected for the composite estimate.

over the tree. Figure 2 illustrates this via a simulation. Branch lengths on 1000 ultrametric and fully-unbalanced (i.e., comb-like) 32-taxon trees were simulated based on a pure birth model for clade growth (Harding, 1971). This is a common process for generating divergence times on trees, with the useful property that the waiting times between successive branching events are drawn from a negative exponential distribution with parameter n , where n is the number of extant taxa at any time (Nee *et al.*, 1992; Nee, 2001). Relative waiting times (and so relative branch lengths) can therefore be simulated simply as $t = -\ln(p) / n$, where p is a uniformly distributed random number between 0 and 1 that represents the uniform distribution of probabilities. Although the trees so generated are all the same size and shape, they differ in their total depth as a result of the stochastic nature of the birth process.

Figures 2a–c depict different calibration scenarios for these simulated trees. In Figure 2a, all 1000 trees were calibrated on the root, forcing them all to have the same total depth. The graph plots the median depth over the sets of equivalent estimated nodes (i.e., the most recent split, the second-most, the third-most, through to the root) as the x -axis and the coefficient of

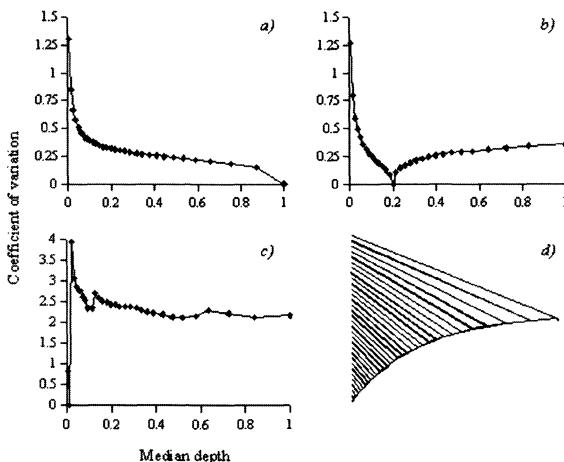


Figure 2. Simulated calibration scenarios: a) calibration on the root, b) calibration on an intermediate node, and c) calibration on a recent node. Each data point represents a set of equivalent nodes over 1000 comb-like, ultrametric trees. For instance, the rightmost point represents a set of a thousand roots, whereas the leftmost point represents the set of nodes that splits the most recent pair of sister species. Median depth over each set is plotted on the horizontal axis such that values of 0 and 1 correspond with the tips and the root, respectively. On the vertical axis, the coefficient of variation over each set is given, give the following calibration scenarios An example of a 32-taxon ultrametric tree with branch lengths simulated under a Yule model, such as the trees used in these calibration scenarios, is given in (d). Its orientation is identical to the data points in (a)–(c) (i.e., with the oldest nodes on the right and the newest on the left).

variation over each of these sets as the y -axis. The data point with the largest depth (i.e., the root) had a coefficient of variation of zero because it was used as the calibration point. As we moved away from the calibration point (i.e., leftward), the coefficient of variation increased because of the cumulative effect of the randomness propagating through the tree. Figure 2b shows how the coefficient of variation behaves when a node of intermediate depth was chosen as calibration point. Once again, the coefficient of variation increased as we moved away from the calibration point, both when we moved nearer to the tips or to the root. However, the mean coefficient of variation over all nodes was lower (here, 0.309 versus 0.368 when calibrated on the root). Figure 2c shows the behaviour when a recent node was used as calibration point: the mean coefficient of variation over all nodes was the highest of all scenarios (2.324). This is probably because of the Central Limit Theorem: the depth of the first split, unlike all that follow, is not the result of the sum of a series of draws from the exponential distribution, but rather of a single draw, and so the variation over a set of such nodes is accordingly higher than that over any set of deeper nodes. Thus, constraining a set of these first,

more variable, splits to the same depth will increase the variation over each set of deeper nodes.

In comb-like trees, all nodes are ordered consistently and linearly, and so the trees in our simulation provide a highly simplified and somewhat extreme example of the effect of choosing a single calibration point on the overall variation over all other nodes. Nevertheless, we expect that the same effect will hold for real data sets, albeit to a lesser extent because most real trees are not fully unbalanced.

Because it is desirable to choose a calibration point that minimizes the total variation over node depth estimates, the best choice would be to choose an intermediate node for calibration. However, even if the variation over different estimates is so minimized, it is still likely to be high as a result of outliers caused by, for instance, 1) saturated genes reducing the estimated depth for deeper nodes or 2) genes that give highly discrepant estimates for other reasons such as different strengths and modes of selection along different lineages. In earlier studies where divergence dates were combined in supertrees (e.g., Purvis, 1995), the influence of such outliers was minimized by taking the median instead of the mean over the set of estimates. We do the same here.

From the simulations, it is evident that overall variation can be reduced by choosing an optimal calibration point. However, even if one were to choose the node that is located optimally within the topology of the tree, stochasticity will still propagate through the tree such that nodes located away from the calibration point will be highly variable. By using multiple calibration points located in disparate regions of the tree, we can minimize this effect. This approach has the added merit of including more previously known information on divergence dates.

Consider Figure 1 again. In this example, the trees were calibrated on the shared node 2. All prior information on the other divergence dates is thus disregarded when obviously we should strive to incorporate all available, robust, information in the estimates. We will do this by averaging all divergence date estimates for a given node across all different calibration points for which prior information is available. For instance, if node 1 in Figure 1 would also be used as a calibration point, we would get two data points for node 2: one where it was used as a calibration point as shown in Figure 1c, and one from tree II calibrated on node 1. Similarly, we would get two estimates for node 3 (one from the median of the estimates obtained by calibrating trees III and IV on node 2, and one from tree I calibrated on node 1) as well as for node 1 (one obtained by calibrating tree II on node 2 and one where it is used as a calibration point for trees I and II). We then average over the data points for each the respective nodes and incorporate the results into the supertree. We apply this method below.

2. Methods

2.1 Phylogeny construction

The primate phylogeny we used in this study will be presented in full in a companion article (Vos and Mooers, in prep.), and so we offer only the briefest outline here. We collected 217 source trees from 126 articles published after 1993 and combined these with the data from the primate supertree of Purvis (1995). We then combined all these data sets into one large MRP matrix using RadCon (Thorley and Page, 2000) and used the parsimony ratchet (Nixon, 1999) strategy as implemented in the program PAUPRat (<http://viceroy.eeb.uconn.edu/paupratweb/pauprat.htm>) to search tree space under various models of character state change. Finally, we constructed majority-rule and strict consensus trees over each of the resulting sets of unique optimal trees.

2.2 Molecular data collection

To collect suitable candidate genes for the inference of relative divergence dates we downloaded the Primates section of the NCBI-GenBank Flat File Release 132.0 from <ftp.ncbi.nlm.nih.gov>. We indexed this data set using the standalone BLAST tool formatdb and performed keyword frequency (“grep”) searches to collect genes that were sequenced over a broad taxonomic range. We refined these results using BLAST (Altschul *et al.*, 1990) searches. This yielded 55 candidate genes. We aligned these sequence data sets using Clustal W’s default settings and method (Thompson *et al.*, 1994) and subsequently by hand. We then ran ModelTEST (Posada and Crandall, 1998) on each data set using the likelihood-ratio test statistic $d = -2 \log L$ to identify the appropriate nucleotide substitution model from a nested set.

Subsequently, we tested whether the molecular clock could be rejected using the same statistical approach, but with a liberal alpha for rejection of 0.001. We chose this alpha level for two reasons. First, given that the likelihood-ratio test for rate constancy is a test of significance, the usual alpha level of 0.05 will reject the clock by chance alone once in every twenty tests on average, even if all loci behave in a clocklike manner (i.e., a Type I error). Lowering the alpha level reduced this risk and so served as a correction for multiple comparisons. Second, lowering the alpha level to 0.001 allowed us to include data sets that evidently deviate somewhat from rate constancy such that they would have been rejected under the more commonly used level of 0.05.

Because this approach by itself yielded too few data sets, we developed a program that iteratively prunes from the non-clocklike data sets those taxa that are the most divergent from the mean root-to-tip path length, and subsequently tests whether the data set then conforms to the molecular clock. The routine stops once $p > 0.001$. Essentially, this program removes those lineages from a data set within which substitution rates have increased or decreased significantly relative to the average of that data set. Data sets where the program stopped when three taxa remained were discarded because conforming to the molecular clock with so few taxa is essentially meaningless.

Using this approach, which could be described as “gene shopping” followed by “taxon shopping”, 40 loci conformed to the molecular clock. The loci analyzed in this study are listed in Table 1; those that conform to the molecular clock and that we used to obtain divergence dates are indicated by an asterisk.

2.3 Inferring and calibrating divergence dates

We labeled each node in the topology of the supertree by appending a serial number — and, to remain compliant with the NEXUS format (Maddison *et al.*, 1997), the word “node” — to each closing bracket of the tree description. The result is similar to the labeling on the tree in Figure 1a. For each aligned clock-like sequence data sets, we then pruned all taxa that were absent in that data set from the supertree so as to obtain constraint trees congruent with the consensus supertree, while keeping track of the initial node-labeling scheme. This resulted in a set of trees with labeled nodes like those shown in Figure 1b. The labeling and pruning was done using Perl scripts, which are available from the authors upon request. We then estimated the branch lengths on these constraint trees under the appropriate models using PAUP* (Swofford, 2002). We calculated relative node depths from these branch lengths using the ape package (<http://stat.ethz.ch/R-CRAN/doc/packages/ape.pdf>) for the R program. The routine that calculates these depths visits all labeled nodes and, for each, calculates the path length from that focal node to the tips and writes it to a table. Because the routine does not take all possible paths into consideration, it gives meaningful results only for ultrametric (i.e., clocklike) trees. We then combined the results from the individual genes into a larger table to calibrate these multiple loci on shared nodes. We surveyed the recent literature for estimates of the timing of major, uncontested splits in the evolutionary history of the primates that could function as calibration points (e.g., Gingerich and Uhen, 1994; Adachi and Hasegawa, 1995; Adachi and Hasegawa, 1996; Arnason *et al.*, 1996a, b,

Table 1. Loci used in this study. The abbreviated models are the following: HKY85: Hasegawa, Kishino, Yano (Hasegawa *et al.*, 1985); K80: Kimura two-parameter (Kimura, 1980); GTR: General Time Reversible (Rodríguez *et al.*, 1990; Yang *et al.*, 1994); +Γ: variation in rates among sites modeled using a gamma distribution (Yang, 1996); +I: a proportion of sites modeled as invariant (Hasegawa *et al.*, 1985). The number of taxa after pruning (see text) is given.

Gene	Model	Clock test <i>p</i> -value	No. of taxa
alpha-1,3-Galactosyltransferase	GTR+I	1.09752 x 10 ⁻²³	19
ATP6	GTR+Γ+I	3.72376 x 10 ⁻¹⁰	17
ATP7A	HKY85+Γ	0.13700341*	7
ATP8	GTR+Γ+I	3.06905 x 10 ⁻¹⁰	17
BRCA1	HKY85+Γ	0.01145129*	7
Calmodulin	HKY85	0.815719539*	6
CCR5	K80+Γ+I	0.00046635	67
CD4	GTR+Γ	0.00189824*	22
COII	GTR+Γ+I	4.56 x 10 ⁻⁸	57
CXCR4	HKY85+Γ+I	8.91 x 10 ⁻⁵	42
DRD4	HKY85+Γ	0.07035867*	14
FUT1	HKY85+Γ	7.3035 x 10 ⁻¹⁴⁹	32
Gamma1 globin	HKY85+Γ	0.0093968*	13
G6PD	HKY85+Γ	0.00121083*	23
IL-2	HKY85	0.92550696*	11
IL-3	GTR+I	0.020268243*	4
IL-4	GTR	0.13912934*	8
IL-6	HKY85	9.88 x 10 ⁻¹⁴	8
IL-10	HKY85	0.16400763*	9
IL-16	GTR+I	0.08135042*	7
Interferon gamma	HKY85+Γ	0.19943861*	13
IRBP (intron 1)	K80+Γ	0.00010886	37
IRBP (partial cds)	HKY85+Γ	0.01551987*	23
LZM	HKY85+Γ	0.000427075	17
nd1	GTR+Γ+I	0.36658552*	12
ND2	GTR+Γ+I	0.00033689	13
ND3	GTR+Γ+I	0.13023191*	36
ND4L	GTR+Γ+I	0.25098333*	45
ND5	GTR+Γ+I	0.00024931	27
ND6	GTR+Γ+I	0.26791855*	12
NRAMP1	HKY85	0.13333399*	14
PLCB4	GTR	0.85023707*	7
PNOC	GTR+Γ	0.28885825*	7
SRY	HKY85+Γ	0.01145427*	59
TSPY	HKY85+Γ	0.00116896*	41
tRNA-ala	GTR+Γ	0.101676857*	10
tRNA-arg	HKY85+Γ+I	0.1082015*	36
tRNA-asn	HKY85+Γ	0.000246114	10
tRNA-asp	HKY85+Γ	0.054571469*	10
tRNA-cys	GTR+Γ	0.597145544*	10

Table 1. Continued.

Gene	Model	Clock test <i>p</i> -value	No. of taxa
tRNA-gln	HKY85+Γ	0.406863018*	9
tRNA-glu	GTR+Γ	0.006563358*	10
tRNA-gly	HKY85+Γ	0.005765017*	30
tRNA-ile	GTR+I	0.003574332*	10
tRNA-lys	HKY85+Γ	0.007495502*	10
tRNA-met	GTR+I	0.552294012*	10
tRNA-phe	HKY85	4.28761 x 10 ⁻⁹	12
tRNA-pro	GTR+Γ	0.22147066*	11
tRNA-thr	GTR+I	0.010557853*	12
tRNA-trp	HKY85+Γ	7.52579 x 10 ⁻⁶	10
tRNA-tyr	GTR+Γ	0.012303697*	10
tRNA-val	HKY85+Γ	0.048727903*	26
ZFX	HKY85+Γ+I	0.0130834*	18
ZFY	GTR+Γ	0.00138935*	13
vWF	HKY85+Γ	0.031415638*	17

1998, 2000; Easteal and Herbert, 1997; Porter *et al.*, 1997; Yoder, 1997; Goodman *et al.*, 1998; Kumar and Hedges, 1998; Stauffer *et al.*, 2001; Nei and Glazko, 2002).

3. Results

The majority-rule consensus tree that we dated was based on a search using irreversible character-state changes, and had a resolution of 0.917 (over 15 242 unique optimal trees), and a consistency index of 0.82.

Figure 3 presents the relationship between the median depth of a set of equivalent estimated nodes and coefficient of variation over that set under three calibration scenarios. Figure 3a depicts the variation over the divergence dates if the depths were calibrated on the split between *Homo* and *Pan*, which is a recent split in the context of primate phylogeny. The total variation was highest under this scenario (mean coefficient of variation, CV = 0.622). Variation was lowered when all node depths were calibrated on the root (mean CV = 0.513; Figure 3c). The coefficient of variation was lowest when the split between the Colobinae and Cercopithecinae was used for calibration (mean CV = 0.345; Figure 3b). The results shown in Figure 3 demonstrate that the actual data behaved as we assumed from the results of our simulations: the lowest overall variation was obtained by calibrating on a node of intermediate depth, whereas recent nodes used as calibration points led to the highest variation. Note that the comparison is not exact for several reasons: first, different numbers of genes were common to each calibration;

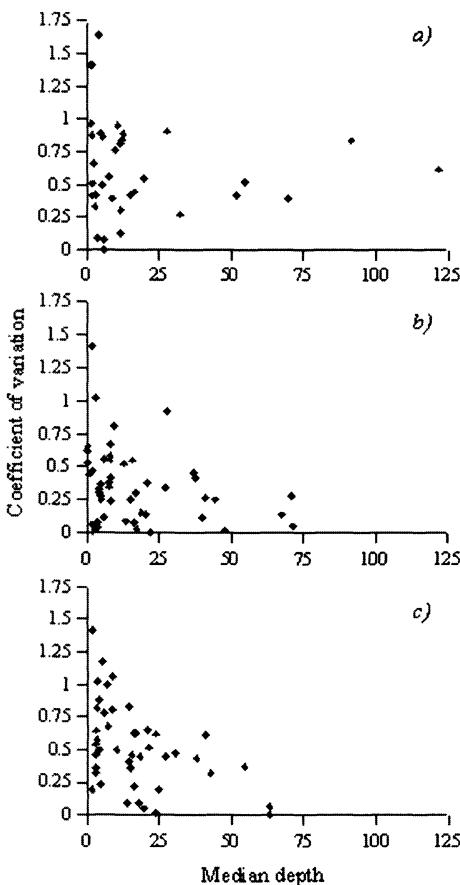


Figure 3. Three calibration scenarios: a) calibration on the split between *Homo sapiens* and *Pan* (the calibration point lies at a depth of 6 MYA); b) calibration on the split between the Cercopithecinae and the Colobinae (22.2 MYA); and c) calibration on the root (63 MYA).

second, the topology of the supertree is not comb-like; and finally, the model used in our simulations was a simplified approximation of the actual process of clade growth (of which a molecular phylogeny is again an approximation).

The depths of the calibration points used in Figure 3 were obtained by taking the median over the estimates we found in a search through the recent literature (Table 2). These previously published dates were obtained through a variety of methods and data sources: from fossils (Goodman *et al.*, 1998); from a coalescence model for species diversity (Gingerich and Uhen, 1994); from maximum likelihood estimates using mtDNA calibrated on divergences

Table 2. Recently published estimates of dates for major primate splits. 1 = apes-Old World monkeys; 2 = *Homo-Pan*; 3 = (*Homo, Pan*)-*Gorilla*; 4 = ((*Homo, Pan*),*Gorilla*)-*Pongo*; 5 = great apes-gibbons; 6 = Old World monkeys-New World monkeys; 7 = root; 8 = lemurs-lorisiforms; 9 = Colobinae-Cercopithecinae. All ages are in millions of years ago.

	1	2	3	4	5	6	7	8	9
Nei and Glazko (2002)	23	6	7			33			
Stauffer <i>et al.</i> (2001)	23	5.4	6.4	11	15				
Gingerich and Uhen (1994)						63			
Yoder (1997)							54		
Arnason <i>et al.</i> (1998)	50					60	80		30
Porter <i>et al.</i> (1997)	25								
Goodman <i>et al.</i> (1998)						38			
Adachi and Hasegawa (1995)		4		16					
Easteal and Herbert (1997)				8.5					
Kumar and Hedges (1998)	23.3	5.5	6.7	8.2	14.6	47.6			
Arnason <i>et al.</i> (1996b)		6.1							
Adachi and Hasegawa (1996)		4.3							
Arnason <i>et al.</i> (2000)	13	16	30	35	70				
Arnason <i>et al.</i> (1996a)	10.4	14.2	19.2	32.4					
Purvis (1995)	27.5	7.0	8.3	14.5	18.2	40.5	57.5	45.1	14.4
Median of published studies	24.1	6	7.6	14.5	18.2	44.0	63	49.5	22.2
Present estimates	32.8	5.9	6.3	15.2	18.8	49.8	77.5	51.6	16.8

outside the order (Arnason *et al.*, 1996a; Arnason *et al.*, 1998), inside the order (Adachi and Hasegawa, 1995, 1996; Yoder, 1997), or calibrated on geological data (Arnason *et al.*, 1996b; Stauffer *et al.*, 2001) or using the method of Li *et al.* (1987; Arnason *et al.*, 2000); from nuclear sequences calibrated on nodes outside (Easteal and Herbert, 1997; Kumar and Hedges, 1998) or inside the order (Porter *et al.*, 1997); from amino acid sequences calibrated inside and outside the order (Nei and Glazko, 2002); and using the mixed fossil and rescaled phylogenies technique outlined earlier (Purvis, 1995). The estimates, all in millions of years ago (MYA), are listed in Table 2. We calibrated our data on the median values over these estimates and averaged over the nine resulting sets of estimates (i.e., one for each calibration point), some of the results of which are listed in the bottom row of Table 2. Figure 4 presents date estimates for the same nine splits we found using our method by calibrating trees first on each of these published estimates in turn and then averaging the results.

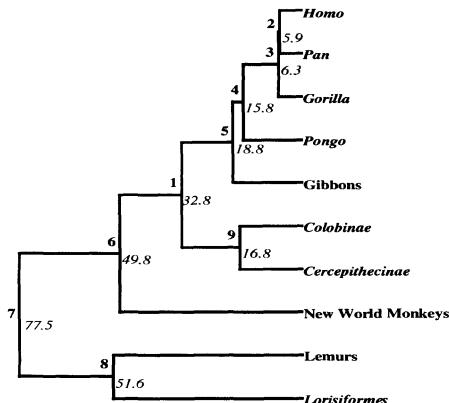


Figure 4. Selected dates of primate divergences using the methods outlined in the text. Numbers above nodes are from Table 2; numbers in front of nodes are divergence dates in MYA.

4. Discussion

The divergence dates estimated using the method described here generally fit well with previously published estimates from different sources (see the examples in Table 2). The correlation between dates estimated here and those for the equivalent nodes in the only other large-scale study (that of Purvis, 1995) is strong (Figure 5). Note that the topology of our supertree is different from that of Purvis in this comparison, and so we compared only those nodes that were unambiguously equivalent (the subtrees descending from these nodes could be different, however). The comparison is therefore not exact, and any differences observed could still be a result of different methods, different topologies, or both.

Compared with the date estimates in Purvis (1995), the estimates presented in this paper were increasingly older with their depth in the tree. We suspect that this is a result of a trend in primate phylogenetics that can be ascribed to both newly discovered, older fossil finds as well as the use of more sophisticated models of sequence evolution in more recent studies.

One potential weakness of our approach is that we have not been able to cover every node in the supertree with the currently available data. On the most resolved topology, 55% of the nodes had date estimates, with all the missing data concentrated around recent nodes in rarely studied clades. Although the amount of sequence data in public databases is growing rapidly, some way of incorporating more non-clocklike loci would seem desirable, perhaps using methods akin to those pioneered by Sanderson (1997, 2002). Even so, missing data points will probably remain in our tree

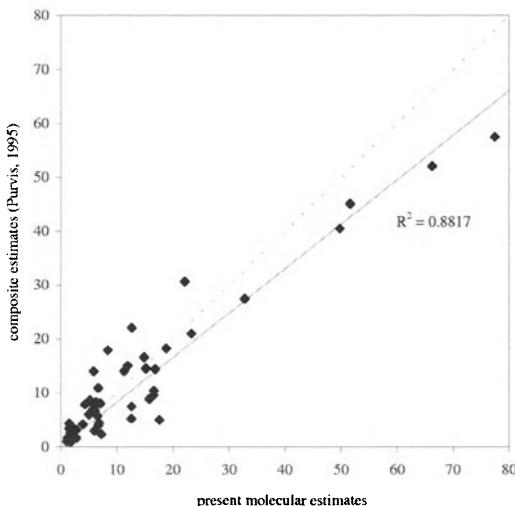


Figure 5. Comparison of previously published composite estimates of divergence dates (from Purvis, 1995) with present estimates. Dotted line indicates 1:1 relationship. See text for further details.

that would have to be interpolated based on models for clade growth such as those used in previous supertree studies (Purvis, 1995; Bininda-Emonds *et al.*, 1999).

More comparisons of our approach with that of Bininda-Emonds *et al.* (1999) will be necessary, as will further exploration of the relative power of this hybrid MRP + model-based method and traditional tree-building algorithms that consider the genetic data directly, incorporate multiple genes and multiple models, and, most dauntingly, mixed clock and nonclock scenarios for different data partitions. This, however, is for the future.

Acknowledgements

We would like to thank Andy Purvis for kindly providing the source data used for his primate supertree research; Vincent Nijman and Eva Chrostowski for assistance with data collection; the members of FAB-lab and Eirikur Palsson for valuable input; Olaf Bininda-Emonds for inviting us to contribute, for his patience, and for his keen editing; and Paul-Michael Agapow and Kate Jones for in-depth reviews of the manuscript.

References

- ADACHI, J. AND HASEGAWA, M. 1995. Improved dating of the human/chimpanzee separation in the mitochondrial DNA tree: heterogeneity among amino acid sites. *Journal of Molecular Evolution* 40:622–628.
- ADACHI, J. AND HASEGAWA, M. 1996. Tempo and mode of synonymous substitutions in mitochondrial DNA of Primates. *Molecular Biology and Evolution* 13:200–208.
- ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W., AND LIPMAN, D. J. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215:403–410.
- ARCHIBALD, J. D. 1999. Molecular dates and the mammalian radiation. *Trends in Ecology and Evolution* 14:278–278.
- ARNASON, U., GULLBERG, A., BURGUETE, A. S., AND JANKE, A. 2000. Molecular estimates of primate divergences and new hypotheses for primate dispersal and the origin of modern humans. *Hereditas* 133:217–228.
- ARNASON, U., GULLBERG, A., AND JANKE, A. 1998. Molecular timing of primate divergences as estimated by two nonprimate calibration points. *Journal of Molecular Evolution* 47:718–727.
- ARNASON, U., GULLBERG, A., JANKE, A., AND XU, X. 1996a. Pattern and timing of evolutionary divergences among hominoids based on analyses of complete mtDNAs. *Journal of Molecular Evolution* 43:650–661.
- ARNASON, U., XU, X. F., GULLBERG, A., AND GRAUR, D. 1996b. The “*Phoca* standard”: An external molecular reference for calibrating recent evolutionary divergences. *Journal of Molecular Evolution* 43:41–45.
- BAUM, B. R. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon* 41:3–10.
- BININDA-EMONDS, O. R. P., GITTLEMAN, J. L. AND PURVIS, A. 1999. Building large trees by combining phylogenetic information: a complete phylogeny of the extant Carnivora (Mammalia). *Biological Reviews* 74:143–175.
- BININDA-EMONDS, O. R. P., GITTLEMAN, J. L., AND STEEL, M. A. 2002. The (super)tree of life: procedures, problems and prospects. *Annual Review of Ecology and Systematics* 33:265–289.
- BROMHAM, L. D., RAMBAUT, A., FORTEY, R., COOPER, A. AND PENNY, D. 1998. Testing the Cambrian explosion hypothesis by using a molecular dating technique. *Proceedings of the National Academy of Sciences of the United States of America* 95:12386–12389.
- BROMHAM, L. D. AND HENDY, M. D. 2000. Can fast early rates reconcile molecular dates with the Cambrian explosions? *Proceedings of the Royal Society of London B* 267:1041–1047.
- BRYANT, D., SEMPLE, C., AND STEEL, M. 2004. Supertree methods for ancestral divergence dates and other applications. In O. R. P. Bininda-Emonds (ed). *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 129–150. Kluwer Academic, Dordrecht, the Netherlands.
- COLLESS, D. H. 1980. Congruence between morphometric and allozyme data for *Menidia* species: a reappraisal. *Systematic Zoology* 29:288–299.
- EASTEAL, S. AND HERBERT, G. 1997. Molecular evidence from the nuclear genome for the time frame of human evolution. *Journal of Molecular Evolution* 44(Suppl. 1):S121–S132.
- FELSENSTEIN, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* 17:368–376.
- FORTEY, R. A., BRIGGS, D. E. G., AND WILLS, M. A. 1996. The Cambrian evolutionary “explosion”: decoupling cladogenesis from morphological disparity. *Biological Journal of the Linnean Society* 57:13–33.

- GINGERICH, P. D. AND UHEN, M. D. 1994. Time of origin of primates. *Journal of Human Evolution* 27:443–445.
- GITTLEMAN, J. L., JONES, K. E., AND PRICE, S. A. 2004. Supertrees: using complete phylogenies in comparative biology. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 439–460. Kluwer Academic, Dordrecht, the Netherlands.
- GOODMAN, M., PORTER, C. A., CZELUSNIAK, J., PAGE, S. L., SCHNEIDER, H., SHOSHANI, J., GUNNELL, G. F., AND GROVES, C. 1998. Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence. *Molecular Phylogenetics and Evolution* 9:585–598.
- GOULD, S. J. 1989. *Wonderful Life*. Norton, New York.
- HARDING, E. F. 1971. The probabilities of rooted tree shapes generated by random bifurcation. *Advanced Applied Probability* 3:44–77.
- HASEGAWA, M., KISHINO, H., AND YANO, T.-A. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* 22:160–174.
- JONES, K. E., PURVIS, A., MACLARNON, A., BININDA-EMMONDS, O. R. P., AND SIMMONS, N. B. 2002. A phylogenetic supertree of the bats (Mammalia: Chiroptera). *Biological Reviews* 77:223–259.
- KIMURA, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* 16:111–120.
- KLUGE, A. AND FARRIS, S. 1969. Quantitative phyletics and the evolution of anurans. *Systematic Zoology* 18:1–32.
- KUMAR, S. AND HEDGES, S. B. 1998. A molecular timescale for vertebrate evolution. *Nature* 392:917–920.
- LEE, M. S. Y. 1999. Molecular clock calibrations and metazoan divergence dates. *Journal of Molecular Evolution* 49:385–391.
- LI W.-H., WOLFE, K. H., SOUDIS, J., AND SHARP, P. 1987. Reconstruction of phylogenetic trees and estimation of divergence times under nonconstant rates of evolution. *Cold Spring Harbor Symposia on Quantitative Biology* 52:847–856.
- LIPPS, J. H. AND SIGNOR, P. W. 1992. *Origin and Early Evolution of Metazoa*. Plenum, New York.
- LIU, F.-G. R., MIYAMOTO, M. M., FREIRE, N. P., ONG, P. Q., TENNANT, M. R., YOUNG, T. S., AND GUGEL, K. F. 2001. Molecular and morphological supertrees for eutherian (placental) mammals. *Science* 291:1786–1789.
- MADDISON, D. R., SWOFFORD, D. L., AND MADDISON, W. P. 1997. NEXUS: an extensible file format for systematic information. *Systematic Biology* 46:590–621.
- MOOERS, A.O., HEARD, S. B., AND E. CHROSTOWSKI, E. In press. Evolutionary heritage as a metric for conservation. In A. Purvis, T. L. Brooks, and J. L. Gittleman (eds), *Phylogeny and Conservation*. Oxford University Press, Oxford.
- MOORE, B. R., CHAN, K. M. A., AND DONOGHUE, M. J. 2004. Detecting diversification rate variation in supertrees. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 487–533. Kluwer Academic, Dordrecht, the Netherlands.
- NEE, S. 2001. Inferring speciation rates from phylogenies. *Evolution* 55:661–668.
- NEE, S., MOOERS, A. Ø., AND HARVEY, P. H. 1992. Tempo and mode of evolution revealed from molecular phylogenies. *Proceedings of the National Academy of Sciences of the United States of America* 89:8322–8326.

- NEI, M. AND GLAZKO, G. V. 2002. Estimation of divergence times for a few mammalian and several primate species. *Journal of Heredity* 93:157–164.
- NIXON, K. 1999. The parsimony ratchet, a new method for rapid parsimony analysis. *Cladistics* 15:407–414.
- NOVACEK M. J. AND WHEELER, Q. D. 1992. Introduction: extinct taxa. In: Novacek M. J. and Q. D. Wheeler (eds), *Extinction and Phylogeny*: New York: Columbia University Press, 1–16.
- PISANI, D., YATES, A. M., LANGER, M. C., AND BENTON, M. J. 2002. A genus-level supertree of the Dinosauria. *Proceedings of the Royal Society of London B* 269:915–921.
- PORTER, C. A., PAGE, S. L., CZELUSNIAK, J., SCHNEIDER, H., SCHNEIDER, M. P. C., SAMPAIO, I., AND GOODMAN, M. 1997. Phylogeny and evolution of selected primates as determined by sequences of the ϵ -globin locus and 5' flanking regions. *International Journal of Primatology* 18:261–295.
- POSADA, D. AND CRANDALL, K. A. 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics* 14:817–818.
- PURVIS, A. 1995. A composite estimate of primate phylogeny. *Philosophical Transactions of the Royal Society of London B* 348:405–421.
- PURVIS, A., NEE, S. AND HARVEY, P. H. 1995. Macroevolutionary inferences from primate phylogeny. *Proceedings of the Royal Society of London B* 260:329–333.
- RAGAN, M. A. 1992. Phylogenetic inference based on matrix representation of trees. *Molecular Phylogenetics and Evolution* 1:53–58.
- RIDLEY, M. 1996. *Evolution*, 2nd edition. Blackwell Science, Inc., Cambridge, Massachusetts.
- RODRÍGUEZ, F., OLIVER, J. L., MARÍN, A., AND MEDINA, J. R. 1990. The general stochastic model of nucleotide substitution. *Journal of Theoretic Biology* 142:485–501.
- SALAMIN, N., HODKINSON, T. R., AND SAVOLAINEN, V. 2002. Building supertrees: an empirical assessment using the grass family (Poaceae). *Systematic Biology* 51:112–126.
- SANDERSON, M. J. 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. *Molecular Biology and Evolution* 14:1218–1231.
- SANDERSON, M. J. 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Molecular Biology and Evolution* 19:101–109.
- STAUFFER, R. L., WALKER, A., RYDER, O. A., LYONS-WEILER, M., AND HEDGES, S. B. 2001. Human and ape molecular clocks and constraints on paleontological hypotheses. *Journal of Heredity* 92:469–474.
- SWOFFORD, D. L. 2002. *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Version 4. Sinauer, Sunderland, Massachusetts.
- THOMPSON, J. D., HIGGINS, D. G., AND GIBSON T. J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Research* 22:4673–4680.
- THORLEY, J. L. AND PAGE, R. D. M. 2000. RadCon: phylogenetic tree comparison and consensus. *Bioinformatics* 16:486–487.
- WILSON, D. E. AND REEDER, D. M. (eds). 1993. *Mammal Species of the World*. Smithsonian Institution Press, Washington DC.
- WOJCIECHOWSKI, M. F., SANDERSON, M. J., STEEL, K. P., AND LISTON, A. 2000. Molecular phylogeny of the “temperate herbaceous tribes” of papilionoid legumes: a supertree approach. In P. Herendeen and A. Bruneau (eds), *Advances in Legume Systematics* 9:277–298. Royal Botanic Garden, Kew.
- XUN, G. 1998. Early metazoan divergence was about 830 million years ago. *Journal of Molecular Evolution* 47:369–371.

- YANG, Z., GOLDMAN, N., AND FRIDAY, A. 1994. Comparison of models for nucleotide substitution used in maximum likelihood phylogenetic estimation. *Molecular Biology and Evolution* 11:316–324.
- YANG, Z. 1996. Among-site variation and its impact on phylogenetic analyses. *Trends in Ecology and Evolution* 11:367–371.
- YODER, A. D. 1997. Back to the future: a synthesis of strepsirrhine systematics. *Evolutionary Anthropology: Issues, News, and Reviews* 6:11–22.
- ZUCKERCANDL, E. AND PAULING, L. 1965. Evolutionary divergence and convergence in proteins. In V. Bryson and H. J. Vogel (eds), pp. 97–165 in *Evolving Genes and Proteins*. Academic Press, New York.

4. A critical look a supertrees

Chapter 14

PERFORMANCE OF SUPERTREE METHODS ON VARIOUS DATA SET DECOMPOSITIONS

Usman Roshan, Bernard M.E. Moret, Tiffani L. Williams and Tandy Warnow

Abstract: Many large-scale phylogenetic reconstruction methods attempt to solve hard optimization problems such as Maximum Parsimony (MP) and Maximum Likelihood (ML), but they are severely limited by the number of taxa that they can handle in a reasonable timeframe. A standard heuristic approach to this problem is the divide-and-conquer strategy: decompose the data set into smaller subsets, solve the subsets (i.e., use MP or ML on each subset to obtain trees), and then combine the solutions to the subsets into a solution for the original data set. This last step — combining given trees into a single tree — is known as supertree construction in computational phylogenetics. The traditional application of supertree methods is to combine existing, published phylogenies into a single phylogeny. Here, we study supertree construction in the context of divide-and-conquer methods for large-scale tree reconstruction. We study several divide-and-conquer approaches and demonstrate experimentally their advantage over the traditional supertree technique of Matrix Representation with Parsimony (MRP), and over global heuristics such as the parsimony ratchet. For the ten large biological data sets under investigation, our study shows that the techniques used for dividing the data set into subproblems as well as those used for merging them into a single solution influence the quality of the supertree construction strongly. In most cases, our merging technique — the Strict Consensus Merger — outperformed MRP with respect to MP scores and running time. Divide-and-conquer techniques are also a highly competitive alternative to global heuristics such as the parsimony ratchet, especially on the more challenging data sets.

Keywords: disk-covering methods; divide-and-conquer; heuristic search methods; maximum parsimony; parsimony ratchet

1. Introduction

Supertree methods combine smaller, overlapping subtrees into a larger tree. Their traditional application has been to combine existing, published phylogenies on which the community agrees into a tree leaf-labeled by the entire set of species. The most popular supertree method is Matrix Representation with Parsimony (MRP; Baum, 1992; Ragan, 1992), which has been used in several phylogenetic studies (e.g., Purvis, 1995; Bininda-Emonds *et al.*, 1999; Liu *et al.*, 2001; Jones *et al.*, 2002; Mahon, 2004). Bininda-Emonds and colleagues (Bininda-Emonds and Sanderson, 2001; Bininda-Emonds, 2003) have evaluated the behavior of several variants of MRP on small, simulated data sets with respect to topological accuracy.

We study the application of supertree methods in a different context: as part of divide-and-conquer methods that can be used to solve difficult optimization problems such as Maximum Parsimony (MP) and Maximum Likelihood (ML) (Felsenstein, 1981; Foulds and Graham, 1982; Steel, 1994; Hillis *et al.*, 1996). These two problems are sufficiently hard that a biologically acceptable phylogenetic analysis can take a very long time (months, perhaps) to derive. The conjecture we study in this paper is that divide-and-conquer strategies can speed up searches for optimal trees under MP and ML.

A divide-and-conquer method for phylogeny reconstruction operates as follows:

- Step 1: Decompose the data set into smaller, overlapping subsets.
- Step 2: Construct phylogenetic trees on the subsets using the desired “base” phylogenetic reconstruction method.
- Step 3: Merge the subtrees into a single (not necessarily fully resolved) tree on the entire data set.
- Step 4: If necessary, refine the resulting tree to produce a binary tree.

Several divide-and-conquer methods have been developed and studied, including quartet-based methods, of which Quartet Puzzling (Strimmer and von Haeseler, 1996) is the most popular, and the family of Disk-Covering Methods (DCMs) (Huson *et al.*, 1999a, b; Nakhleh *et al.*, 2001; Warnow *et al.*, 2001; Tang and Moret, 2003). In each of these methods, a supertree method (Step 3) is used to combine subtrees into a tree on the entire data set. Supertree methods are thus an integral aspect of a divide-and-conquer strategy, but the other three aspects of such a strategy also affect accuracy and speed. Our study addresses the following questions:

- Should the subtrees used in reconstruction be selected carefully in terms of the subsets they represent or can the subsets be arbitrary so long as some overlap exists among them?
- Given a fixed collection of overlapping subtrees, what is the best method to assemble them into a single supertree?
- How do divide-and-conquer methods fare when compared to “global” approaches such as the heuristic MP searches in PAUP* (Swofford, 2002)?

To investigate the first two questions, we compared methods that differ explicitly in how they decompose the data set and how they merge subtrees into a supertree. We considered two variants in the DCM family (DCM1 and DCM2), plus (as a control) random decompositions; these decompositions were coupled with MRP and/or the Strict Consensus Merger (SCM; the supertree method developed for the DCM family) to merge the resulting subtrees into a single supertree; finally, all combinations of methods were followed by a refinement phase. To ascertain whether divide-and-conquer approaches can outperform “global” approaches to solving MP or ML, we compared the performance of our DCM strategies with the parsimony ratchet (Nixon, 1999), one of the best performing MP heuristics for large data sets.

1.1 Overview of experimental results

We compared these methods on ten biological data sets that contain between 328 and 854 taxa, focusing on how the techniques used for data set decomposition and supertree reconstruction impacted on the running time and the MP score of the result. We found that the DCM2 + SCM method outperformed all other methods on all our data sets. The specific decomposition technique had a significant impact on the MP score of the resulting tree as well as on running time, with DCM2 clearly outperforming random decompositions. Furthermore, we obtained improved MP scores in all decomposition strategies (DCM and random) when the subproblems were large, an observation that impacts on taxon-sampling strategies. The supertree method used to combine subtrees into a single tree on the full data set was also very important. When MRP and SCM were followed by the same resolution technique in Step 4, SCM generally produced better MP scores than MRP. The only exception was for DCM1-based decompositions, but, as our results show, these decompositions are relatively poor and not competitive.

Our study demonstrated that the benefits of a divide-and-conquer technique depend on the properties of the data set. When the data set can be decomposed well by DCM2 into significantly smaller subproblems with

good overlap, DCM2 provided a clear advantage in both running time and MP scores. The advantage was most pronounced for challenging data sets, data sets for which heuristic MP searches take a long time to find a first good solution. We compared DCM2-based approaches with the parsimony ratchet, the best MP global heuristic in our experiments, on two biological data sets: the well-studied 500 *rbcL* DNA data set (Chase *et al.*, 1993; Rice *et al.*, 1997) and a set of 816 bacterial rRNA sequences (Wuyts *et al.*, 2002). The *rbcL* data set decomposes poorly and is not especially challenging for MP heuristics; our study shows that DCM2 provided no improvement over the parsimony ratchet for this problem. (Interestingly, DCM2-Ratchet, which uses the parsimony ratchet as a base method in a DCM2 decomposition, was almost as good as a global ratchet on the *rbcL* data set in spite of the very poor decomposition.) By contrast, the rRNA data set is challenging for MP heuristics, but decomposes well; our study showed that DCM2 clearly improved on the parsimony ratchet for this problem.

1.2 Comparison with previous work

Bininda-Emonds and colleagues (Bininda-Emonds and Sanderson, 2001; Bininda-Emonds, 2003) studied supertree reconstruction from an experimental point of view, focusing on the MRP method and using small, simulated data sets. Although we also study MRP, our focus is as much on decomposition as it is on supertree reconstruction and so we studied several other methods. Moreover, our testing used biological data sets rather than simulated ones, thereby forcing us to use MP scores as our measure of accuracy (because the true trees for these data sets are not known). Finally, we focused on large data sets (limited in this study to data sets with less than 1000 taxa as a result of the dearth of larger published data sets) because these are the data sets where a divide-and-conquer methodology should have the largest impact.

Some of the earliest divide-and-conquer methods are quartet-based methods such as Quartet Puzzling (Strimmer and von Haeseler, 1996), Short Quartet methods (Erdős *et al.*, 1997), and Quartet Cleaning (Berry *et al.*, 1999). Quartet methods are at one extreme of divide-and-conquer methods because they decompose the data sets into the smallest possible subsets for which nontrivial trees exist — subsets of just four taxa each. Quartet-based methods cannot use either of the two supertree reconstruction techniques we study here profitably: MRP is too expensive given the tiny trees and SCM will usually return a totally unresolved tree because too many quartets will be in conflict. In an earlier study (St. John *et al.*, 2001), we compared various quartet-based methods and the fast and simple neighbor-joining method (NJ; Saitou and Nei, 1987) on simulated data. Quartet Puzzling,

which merges quartet trees using a greedy heuristic, clearly dominated the other quartet-based methods, but was much slower and clearly less accurate than NJ. These results suggest that decomposition into tiny subsets is not profitable. Other published divide-and-conquer methods include Compartmentalization (Mishler, 1994), which is not described fully and so cannot be implemented, and a strategy used to analyze a biological data set (Olsen *et al.*, 1994), where again the decomposition and merging steps are not described well enough to enable us to implement and test the strategy.

2. Divide-and-conquer reconstruction methods

Recall that a divide-and-conquer method uses four basic steps to construct a supertree from a given data set S of n sequences:

- Step 1: The set S is divided into overlapping subsets, S_1, S_2, \dots, S_p .
- Step 2: A tree T_i is constructed on each subset S_i by some “base method” (e.g., a heuristic MP or ML search).
- Step 3: A supertree method is applied to the set of subtrees $\{T_i : i = 1, 2, \dots, p\}$ to obtain a tree T on the full data set.
- Step 4: If the tree T is not fully resolved (i.e., if it contains nodes of degree greater than three), a refinement technique is used to produce a binary tree refining T that optimizes the chosen criterion.

Steps 2 and 4 are the same in all our algorithms except for our study of global heuristics versus divide-and-conquer methods in Section 5. We use a slow heuristic MP search as the “base method” to construct the subtrees, but a fast heuristic MP search to refine the merged supertree into a binary tree. Thus, our methods differ only in how they implement Steps 1 and 3. Sections 2.1 and 2.3 describe the techniques used for data decomposition and subtree merging, respectively. A summary of all of the supertree methods used in our study is given in Section 2.5.

2.1 Data decomposition

2.1.1 DCM-based decomposition

DCMs (Huson *et al.*, 1999a, b; Nakhleh *et al.*, 2001; Warnow *et al.*, 2001) are meta-methods for phylogenetic reconstruction: they operate in conjunction with a “base method” such as an MP heuristic or NJ. DCMs

decompose the input set into smaller overlapping sets on which subtrees are computed using the specified base method. They have a dual goal: improved accuracy and better speed. Because the subsets have smaller diameters (maximum pairwise distances) than the original data set, they are less likely to cause accuracy problems, and because the possibly expensive base methods have to solve only small subsets, the overall algorithm runs faster. One goal can be stressed at the expense of the other; thus, there are several DCMs, each of which was designed for use with a particular base method.

The first DCM, DCM1 (Huson *et al.*, 1999a), was designed for methods such as NJ, where large pairwise distances affect the topological accuracy negatively. DCM1 thus attempts to minimize the evolutionary diameter of each subproblem: it produces many subproblems each with small diameter, but does not control the overlap between the subproblems. Earlier studies we conducted (and confirmed here) showed that DCM1 does not work particularly well with heuristic MP as a base method. Therefore, we developed DCM2 (Huson *et al.*, 1999b), which produces a small number of subproblems (two or three is typical in experiments presented here), all of which share one subset of taxa and are otherwise disjoint. Thus, DCM2 controls the overlap pattern tightly, but does not attempt to control the diameter of each subset directly, thereby producing larger disks than DCM1.

The input to both DCM1 and DCM2 is a set $S = \{s_1, \dots, s_n\}$ of n taxa (typically, aligned biomolecular sequences); an $n \times n$ matrix, $D = (d_{ij})$, containing an estimate of the pairwise distances between the taxa; and a *threshold*, a particular $q \in \{d_{ij}\}$. Both methods start by computing a *threshold graph*, $G(d, q)$, defined as follows:

- The vertices of $G(d, q)$ are the taxa, s_1, s_2, \dots, s_n .
- The edges of $G(d, q)$ are those pairs (s_i, s_j) obeying $d_{ij} \leq q$.

The graph is then minimally *triangulated*; in other words, edges are added to the graph until every cycle of length at least four has a chord (an edge connecting two non-consecutive vertices on the cycle) (Buneman, 1974; Golumbic, 1980), while attempting to minimize the weight of the largest edge added. Obtaining an optimal triangulation of a graph is NP-hard in general (Bodlaender *et al.*, 1992), but threshold graphs are usually triangulated or close to it (Huson *et al.*, 1999a), and our experience shows that even the simple greedy heuristic produces triangulations that do not have very long edges. We triangulate the threshold graph because triangulated graphs have many computationally useful properties, notably:

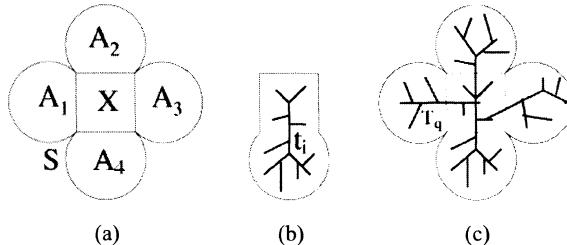


Figure 1. The three steps of Phase I in DCM2: a) compute clique separator X for set S in threshold graph $G(d, q)$, producing subproblems $A_1 \cup X, A_2 \cup X, \dots, A_r \cup X$; b) compute tree t_i for each subproblem $A_i \cup X$; and c) merge computed subtrees to obtain tree T_q for set S .

- they have a linear number of maximal cliques (cliques that cannot be increased by adding a vertex) and these cliques can be computed in polynomial time, and
- their minimal vertex separators (minimal connected subgraphs whose removal breaks the graphs into disconnected pieces) are maximal cliques.

(By contrast, these two problems are NP-hard for general graphs.) Thus, the next step in DCMs is to compute the maximal cliques. At this point, DCM1 is done and simply returns these cliques as the subproblems in the decomposition; these cliques have low diameter by construction. DCM2 scans through the cliques to find one clique X that minimizes $\max_i |X \cup A_i|$, where the A_i are the pieces into which the graph is broken upon removal of X ; it then returns the subsets $X \cup A_i$ as the subproblems in the decomposition. Note that these subproblems have a unique common intersection, but that their diameter can be much larger (because of the addition of the separator X) than that of a subproblem generated by DCM1. Figure 1 shows a symbolic representation of the DCM2 decomposition. We have proved elsewhere that, so long as the subtrees are inferred correctly and the subproblems are large enough, the SCM technique applied to the subtrees will produce the true tree (Huson *et al.*, 1999a). These theorems have ramifications for both DCM1- and DCM2-based strategies, but especially for DCM1 combined with distance-based methods; for these combinations, it is possible to prove nice theorems about the sequence length requirements of the resulting methods (Huson *et al.*, 1999a; Warnow *et al.*, 2001).

However, our goal in this study is practical rather than theoretical: we want to develop faster and more accurate phylogenetic search algorithms that perform well in practice. We experimented with different

decompositions to determine which ones produced the best empirical results, so that improved MP scores are obtained faster. We picked a minimum triangulation to avoid grouping taxa that are evolutionarily distant (which would result in long edges in the triangulated graph). In developing the threshold graph, we needed to choose a threshold q . The smallest useful value for q is d_0 , the smallest possible value for which the threshold graph $G(d, q)$ is connected; the largest possible value is simply $\max\{d_{ij}\}$. (If we applied the algorithm to this largest value, we would not obtain any decomposition into smaller subproblems because the threshold graph would already be a clique.) In our experiments, we looked at ten equally spaced values between d_0 and $d_{10} = \max\{d_{ij}\}$, and ran all tests with values d_0 and d_4 .

2.1.2 Random decomposition

As a control for DCM2, we also considered the effects of decomposing a data set into random overlapping subsets as determined using three parameters: the number x of subproblems, the desired minimum size y of each subproblem, and the desired minimum size z of the pairwise intersection of subsets. Let n be the number of taxa to be distributed among the subsets. The x subsets are populated as follows. First, z taxa are selected randomly and all are placed into each of the subsets. For each subset, we then select an additional $y - z$ taxa randomly from the remaining available taxa (marking the chosen $y - z$ taxa as unavailable). Finally, if any taxa have not yet been placed in any particular subset, we add these taxa randomly to subsets. The resulting decomposition mimics the structure of DCM2 in that it produces subsets with a shared subset, but that are otherwise pairwise disjoint.

2.2 Base methods: heuristic searches for MP

Heuristic searches for MP trees form a basic part of our divide-and-conquer reconstructions in three places: using a base method on subproblems to construct subtrees (Step 2), using MRP to merge subtrees into a supertree (Step 3), and refining the resulting tree into a binary tree (Step 4). The heuristic MP search (HS) of PAUP* v4.0b10 (Swofford, 2002) was used for these analyses because the data sets are too large for exact optimization. Experiments were performed on simulated data to determine the quality of the HS needed in each stage.

- Fast HS: a fast heuristic search in which we save only one tree starting from one initial random sequence addition ordering. We used the PAUP* commands:

- ```

set criterion=parsimony maxtrees=1 increase=no;
hsearch start=stepwise addseq=random swap=tbr hold=1
nreps=1;

```
- Medium HS: medium heuristic search with ten random sequence addition orderings and 100 saved trees. We used the PAUP\* commands:

```

set criterion=parsimony maxtrees=100 increase=no;
hsearch start=stepwise addseq=random swap=tbr hold=1
nreps=10;
contree all/ strict=yes;

```
  - Slow HS: a slow heuristic search with 100 random sequence addition orderings and 1000 saved trees. We used the PAUP\* commands:

```

set criterion=parsimony maxtrees=100 increase=no;
hsearch start=stepwise addseq=random nreps=100 nchuck=1
chuckscore=1 swap=tbr; set maxtrees=1000 increase=no;
filter best=yes; hsearch start=current swap=tbr hold=1
nchuck=1000 timelimit=3600;
contree all/ strict=yes;

```

We also used the parsimony ratchet (Nixon, 1999) in a PAUP\* implementation written by Olaf Bininda-Emonds (perlRat; available from <http://www.tierzucht.tum.de/Bininda-Emonds/>). The ratchet is a simple and effective heuristic for general optimizing search and works iteratively as follows:

1. Run Fast HS for MP.
2. Select 25% of the sites randomly, set their weights to two and run Fast HS on the perturbed data, starting with the tree from the previous search.
3. Reset the site weights to their original values and run Fast HS starting with the tree from the previous search.
4. Repeat steps two and three as desired.

## 2.3 Merging subtrees

### 2.3.1 Matrix representation parsimony

The MRP approach encodes a set  $T$  of trees as binary characters with missing values (i.e., “partial binary characters”) and then applies some heuristic for MP on the resulting set of sequences. Understanding how MRP

works thus requires understanding the encoding and the interpretation of partial binary characters.

Let  $S$  denote the full set of taxa and let  $T$  be one of the trees in the set  $\mathcal{T}$ ; thus,  $T$  has leaf set  $S_0 \subset S$ . Let  $e$  be an arbitrary edge in  $T$ . Deleting  $e$  from  $T$  partitions the leaves of  $T$  into two sets,  $A$  and  $B$ . Now define a character  $c_e$  on all of  $S$  by setting

$$c_e(s) = \begin{cases} 0 & \text{if } s \in A \\ 1 & \text{if } s \in B \\ ? & \text{otherwise} \end{cases}$$

The set  $C(\mathcal{T}) = \{c_e : \exists T \in \mathcal{T}, e \in E(T)\}$  is the MRP encoding of the set  $\mathcal{T}$  of trees.

Given a set of sequences defined by partial binary characters and a candidate tree  $T$  on the set of sequences, all  $?$ s are replaced by 0 or 1 in such a way as to minimize the parsimony score of the tree. If every subtree in the MRP analysis is accurate (i.e., topologically identical to the true tree induced on its set of leaves), then the true tree is one of the MP trees. Hence, an exact solution to MP will return the true tree as one of the solutions. This observation follows from the fact that the true tree is a “perfect phylogeny” (Bodlaender *et al.*, 1992) for the MRP-encoded set of sequences.

For MRP, we used Slow HS. Because the search can identify more than one tree of lowest score, we returned the strict consensus of all the best trees found (i.e., the most resolved tree that is a common contraction of these best trees).

### 2.3.2 Strict-consensus merger

The SCM combines a set of trees into a single tree. The merging is done pairwise until only one tree is left. The specific order in which the trees are merged matters when the subtrees are defined by a DCM1 decomposition, but is irrelevant when the subtrees are defined by a DCM2 decomposition. Hence, it suffices for DCM2 to describe how SCM operates on two trees (for specifics on how SCM operates with DCM1, see Huson *et al.*, 1999a).

Let  $L(T)$  denote the set of leaves of  $T$ ;  $C(T)$  denote the set of bipartitions of  $T$ ; and  $T_X$ , with  $X \subseteq L(T)$ , denote the tree obtained by restricting the leaf set of  $T$  to  $X$  and suppressing nodes of degree two (see Figure 2). SCM takes two trees  $T_1$  and  $T_2$  and returns a tree  $T_{12}$  on the leaf set  $L(T_1) \cup L(T_2)$  according to the following procedure:

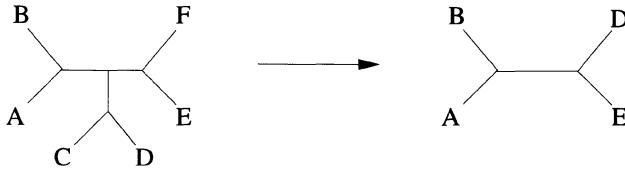


Figure 2. Tree  $T$  restricted to leaf set  $\{A, B, D, E\}$ .

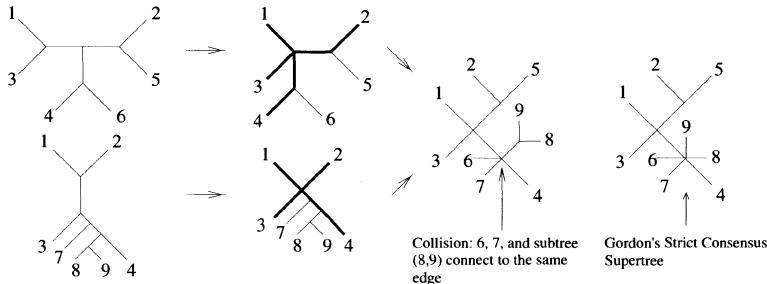


Figure 3. Handling collisions in the SCM and Gordon's Strict Consensus Supertree. The bipartition  $\{\{1, 2, 3, 4, 5, 6, 7\}, \{8, 9\}\}$  is present in the supertree under SCM, but not in Gordon's Strict Consensus Supertree.

- Set  $X = L(T_1) \cap L(T_2)$ .  $X$  is the *backbone* and must satisfy  $|X| \geq 3$ .
- Compute the strict consensus,  $T_X$ , of  $T_1$  and  $T_2$ , each restricted to the leaf set  $X$ .
- Add the remaining taxa from  $T_1$  and  $T_2$  into  $T_X$  to form  $T$ , so as to preserve as much structure as possible. Some piece of each tree  $T_1$  and  $T_2$  may attach onto the same edge of  $T_X$  (causing a *collision*).

Figure 3 illustrates the SCM algorithm on incompatible trees with a collision (i.e., an edge in the backbone to which both trees contribute pieces; the backbone is highlighted with thick edges). SCM handles collisions in the following way. If an edge  $e$  of the backbone has a collision, then we subdivide the edge to produce a new node  $v_e$  to which all contributions will be attached; in each subtree  $T$  contributing to this edge, we identify all pieces of  $T$  that should attach to that edge and attach them directly to  $v_e$ .

The SCM of two trees is very similar to Gordon's Strict Consensus Supertree (SCS; Gordon, 1986). When the trees are compatible (and there is no collision), the two methods produce the same output. However, when there is a collision, the SCS tree can be a strict contraction of the SCM tree because it might contract additional edges located within pieces involved in the collision.

## 2.4 Optimal tree refinement

Merging subtrees into supertrees using MRP or SCM can result in unresolved trees. All steps up to and including the merging step are perhaps seen best as attempts to identify the best-supported edges. Resolving the remaining polytomies (by adding edges) so as to minimize the parsimony score of the resulting tree is the NP-hard *Optimal Tree Refinement (OTR)* problem (Bonet *et al.*, 1998). To “solve” it here, we passed unresolved trees as constraint trees to PAUP\* and used a fast MP heuristic search to produce a resolved tree using the following commands:

```
constraints c1 (monophly) = <the unresolved tree that is used as
constraint>;
set criterion=parsimony maxtrees=1 increase=no;
hsearch start=stepwise addseq=random swap=tbr hold=1 nreps=1
constraints=c1 enforce=yes;
```

## 2.5 Supertree methods studied

By varying the techniques used to decompose the data set into subsets and those used to merge subtrees into supertrees, we can obtain many different divide-and-conquer methods. For each method, a choice of parameters exists. We studied DCM1 and DCM2, each with a supertree construction phase of MRP or SCM, plus random decomposition followed by MRP. (SCM can construct supertrees on an arbitrary set of subtrees, but because the order in which the trees are merged can have a big impact on the resulting supertree, more research is needed to understand how SCM performs with random decompositions.) For DCM1, we used only threshold  $d_0$ , whereas for DCM2 we used both  $d_0$  and  $d_4$ . The methods we tested were:

- $DCM1 + SCM(d_0)$
- $DCM2 + SCM(d_0)$
- $DCM2 + SCM(d_4)$
- $DCM1 + MRP(d_0)$
- $DCM2 + MRP(d_0)$
- $DCM2 + MRP(d_4)$
- $RANDOM + MRP$

## 3. Experimental Methodology

We ran two sets of experiments. The first set was designed to test two conjectures: 1) that careful decomposition of the data set is crucial to the

success of supertree methods and that the DCMs offer such a careful decomposition, and 2) that the SCM developed as part of DCMs is superior to MRP as a supertree-assembly tool. The second set of experiments was designed to test our conjecture that divide-and-conquer methods are a competitive alternative to global heuristics.

We used large biological data sets to test these conjectures. Biological data sets suffer from several disadvantages when used in testing algorithms: 1) we cannot produce “tailored” biological data sets designed to test specific aspects of the reconstruction algorithms; 2) we cannot judge the outcomes on the basis of accuracy (because we do not know the “true” tree), and so must rely instead on substitute criteria such as MP or ML scores; and 3) we cannot use the data sets to predict behavior on other data sets (because we do not know how to relate the specific characteristics of one biological data set to those of another). By contrast, biological data sets offer data with all the biases and peculiarities that are so hard to produce in simulations. Thus the main use of “real-world” data is in spot-checking (Moret, 2002): confirming that predictions made on the basis of simulation results hold for biological data or pinpointing problems with models when the data sets yield incompatible results. In this study, we chose biological data sets for two reasons: 1) we needed them for spot-checking our conjectures, which we derived from large-scale simulation studies that we have conducted already (St. John *et al.*, 2001; Nakhleh *et al.*, 2002; Moret *et al.*, 2002), and 2) no comparable experimental study has been conducted, with existing reports being limited to small biological data sets or to just one larger data set.

Because our conjectures could hold in significant parts of the parameter space, but not everywhere, we studied the effect of various parameter settings. We parameterized the decomposition in terms of subset sizes and mean coverage (i.e., the mean number of subsets in which a taxon appears). Each taxon must appear in at least one subset, but reconstruction requires mean coverage greater than 1 $\times$  (otherwise we would obtain a bush and not a tree). We matched the size and coverage characteristics of random decompositions to our DCM decompositions and studied the variation in parsimony scores as a function of subset sizes or coverage.

### 3.1 The data sets

We obtained ten biological data sets (all of biomolecular sequences) from various sources. Below we give a brief description of each data set, noting the number of sequences, their lengths, and the maximum p-distance (normalized Hamming distance) between any two sequences in the set.

1. 328 ITS RNA sequences (946 sites) from the flowering plant family Asteraceae obtained from the Gutell Lab at the Institute for Cellular and Molecular Biology, The University of Texas at Austin; max p-distance = 0.524.
2. 439 aligned rDNA sequences of eukaryotes (2461 sites) (Goloboff, 1999); max p-distance = 0.649.
3. 476 aligned metazoan DNA sequences (1008 sites) (Goloboff, 1999); max p-distance = 0.445.
4. 500 aligned *rbcL* DNA sequences for flowering plants (759 sites) (Chase *et al.*, 1993; Rice *et al.*, 1997); max p-distance = 0.334.
5. 556 aligned 16S rRNA sequences (2402 sites) for the Spirochaetes class of bacteria (Maidak *et al.*, 2001); max p-distance = 0.310.
6. 567 “three-gene” (*rbcL*, *atpB*, and 18S rDNA) aligned DNA sequences for flowering plants (4592 sites) (Soltis *et al.*, 2000); max p-distance = 0.150.
7. 590 aligned small subunit Archaea rRNA sequences (1962 sites) (Wuyts *et al.*, 2002); max p-distance = 0.382.
8. 695 aligned 16S rRNA sequences (2550 sites) for the Cyanobacteria class of bacteria (Maidak *et al.*, 2001); max p-distance = 0.219.
9. 816 aligned 16S rRNA sequences (1253 sites) for the Bifidobacteriales (132), Acholeplasmataceae (234), and Flexibacteraceae (450) families of bacteria (Wuyts *et al.*, 2002); max p-distance = 0.460.
10. 854 aligned *rbcL* DNA sequences (937 sites) (Goloboff, 1999); max p-distance = 0.390.

Recall that DCM-based approaches require a distance matrix to compute the threshold graph in the first step of their computation, but that the distance matrix does not play any role in the phylogenetic reconstruction beyond this. We used the Kimura two-parameter plus gamma distance-correction formula (Kimura, 1980; Yang, 1993) to compute a distance matrix for each data set using the “default” parameter values of  $\kappa = 2$  and  $\alpha = 1$ . We do not need the model to fit the data particularly well because it affects only the choice of edges in the threshold graph; nevertheless, it is possible that a better distance correction model (such as could be determined using ModelTEST; Posada and Crandall, 1998) would yield better results. In other words, our results with the DCM-based approaches should be regarded as pessimistic; they establish a lower bound, but are subject to further improvement.

### 3.2 Implementation and platforms

Our DCM implementations are a combination of C++ (which uses LEDA 4.3) and Perl scripts; they were written originally by Daniel Huson and

expanded further by us. The random decomposition is also a combination of C++ and Perl scripts and was written by us. To run the MP heuristics used for solving the subproblems, for MRP, and for OTR, we use constrained searches as implemented in PAUP\*.

Our experiments were run on two sets of processors running Debian Linux: the phylofarm cluster of nine dual 500MHz Pentium III processors, and 16 dual 733MHz Pentium III processors that are part of the 132-processor SCOUT cluster. For our running-time analysis, we provide the running time (in seconds) of the following four major steps separately:

- Decomposition: For the DCM-based methods, this includes the running time for computing and triangulating the threshold graph and finding the subproblems. For the random methods, it is the time to form the subproblems.
- Base method: This is the total running time for Slow HS on all the subproblems.
- Merge: This is the running time to merge the subtrees into a supertree using MRP or SCM.
- OTR: This is the cost of running Fast HS with the (unresolved) tree obtained in the previous step as a constraint tree in PAUP\*.

We show only selected data in the following sections; the complete data from our experiments is available on our website at [www.cs.utexas.edu/users/phylo/supertree chapter/](http://www.cs.utexas.edu/users/phylo/supertree/chapter/).

## 4. Results on decompositions

### 4.1 Comparing different DCMs

We begin by examining the six different DCM-based approaches defined in Section 2.5. Figures 4 and 5 show relative MP scores and running times on the ten data sets. The best method (in terms of MP scores) was consistently DCM2 + SCM at either  $d_0$  or  $d_4$ ; the other methods are not nearly as competitive. (MP scores that are lower even by these small percentages are often considered “significant” in phylogenetic analysis.)

Furthermore, SCM is better than MRP at combining subtrees in nearly all cases. (The only exception was for DCM1 decompositions, but these decompositions were relatively poor and clearly not competitive.) Moreover, running times show that MRP was far more expensive than SCM, a matter of

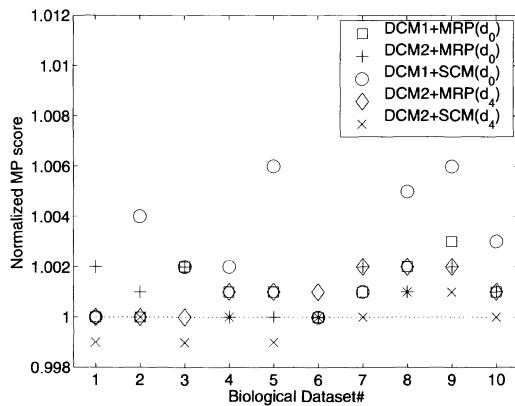


Figure 4. Comparison of the MP scores of DCM-based approaches on ten biological data sets, normalized with respect to the MP score of DCM2 + SCM( $d_0$ ).

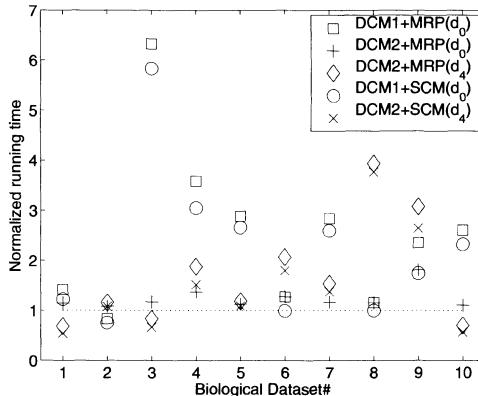


Figure 5. Comparison of the running times of DCM-based approaches on ten biological data sets, normalized with respect to the running time of DCM2 + SCM( $d_0$ ).

hours versus seconds, respectively. Therefore, we focus on DCM2 + SCM because it is clearly the best divide-and-conquer strategy that we tested. Our first task was to determine a suitable threshold. Figures 6 and 7 (using the data of Figures 4 and 5) show that DCM2 + SCM( $d_4$ ) outperformed DCM2 + SCM( $d_0$ ) on most of the ten data sets. This improvement in MP scores as we increased the threshold value is consistent with previous studies (Huson *et al.*, 1999b) and our recent simulation studies (not shown). Note that, as we increased the threshold, the maximum subproblem size increased, but the number of subproblems decreased; hence, the total running time

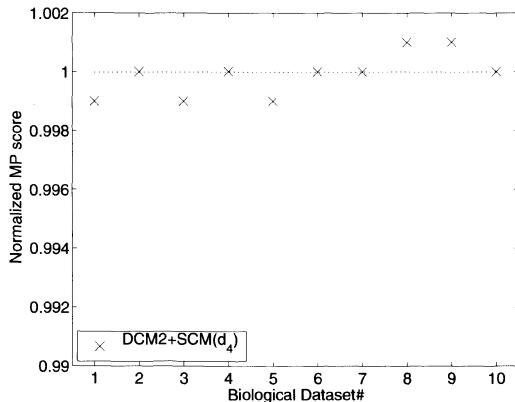


Figure 6. The ratio of the MP scores of DCM2 + SCM( $d_0$ ) to those of DCM2 + SCM( $d_4$ ) on ten biological data sets.

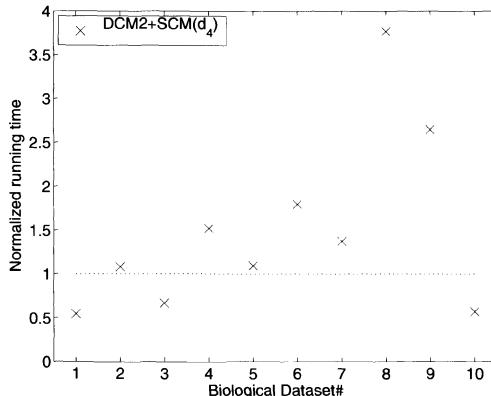


Figure 7. The ratio of the running times of DCM2 + SCM( $d_0$ ) to those of DCM2+SCM( $d_4$ ) on ten biological data sets.

could decrease. For DCM2 + SCM, whether for threshold  $d_0$  or  $d_4$ , by far the most costly aspect of the reconstruction was the time spent in the MP heuristics for reconstructing trees on the subsets and, to a lesser extent, in the OTR phase. By contrast, the DCM and SCM phases were very fast (data not shown, but are available from our website.)

## 4.2 Comparing random decompositions

With random decompositions, we must use the MRP supertree method because SCM is designed specifically for DCMs. Our goal here was to

understand the effect of the (random) decomposition — in particular, the size of subsets and the amount of coverage (i.e., the average number of subsets in which a taxon appears) — on the quality of reconstructions.

We wanted the coverage to run from 2 $\times$  to 5 $\times$  and the size of the subproblems to range from 10% to 90% of the data set. Thus, we chose the number of subproblems to be:

$$\text{number of subproblems} = \text{floor}\left(\frac{\text{coverage total problem}}{\text{subproblem size}}\right)$$

To not bias the subset decomposition further, we set the parameter  $z$  (the minimum overlap size) to 0 and let the pairwise overlap be induced through the number of subproblems and subproblem sizes as chosen above.

We solved the subproblems using Fast HS for MP (see Section 2) and ran MRP using Medium HS for MP (and defined by a time limit of 600 seconds). We examined five values of average subproblem sizes: 10%, 30%, 50%, 70%, and 90% of the data set. For each average subproblem size, we examine coverages of 2 $\times$ , 3 $\times$ , 4 $\times$ , and 5 $\times$ . The detailed results for each of the ten data sets are available on our website. Unsurprisingly, we found that MRP applied to random decomposition did much better with larger subsets and somewhat better with increased coverage, as was also observed by Bininda-Emonds and Sanderson (2001). Furthermore, as the subproblem sizes became larger, the MP scores of MRP on random decompositions slowly approached those of DCM2 + SCM( $d_0$ ).

### 4.3 DCM versus random decompositions

We now explore the relative performance of DCM2 and random decompositions. We ran DCM2 + SCM( $d_0$ ) as a benchmark only, but focus on DCM2 + MRP( $d_0$ ) because we can ensure that MRP is applied to closely comparable decompositions. (We can set the three parameters for random decomposition so as to produce the same number of subsets as DCM2, with closely matched average subset sizes and coverage.) We used Slow HS for MP on the subproblems as well as for MRP and again report the average over five runs for the random decomposition. Figures 8 and 9 plot the ratios of MP scores and running times, respectively, of DCM2 + MRP and of MRP on random decompositions to those of DCM2 + SCM. (MP scores for the trees obtained by each method, along with other details, can be found on our website.) The results clearly indicate that DCM2 + SCM did better in terms of both MP scores and running times than either DCM2 + MRP or MRP on

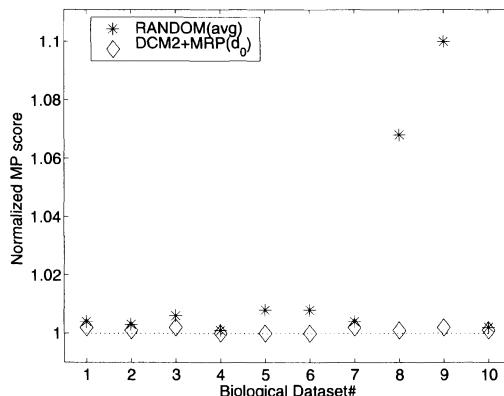


Figure 8. Comparison of MP scores of DCM-based methods and RANDOM (averaged over five runs) normalized with respect to the DCM2 + SCM( $d_0$ ) MP scores on each of the ten biological data sets.

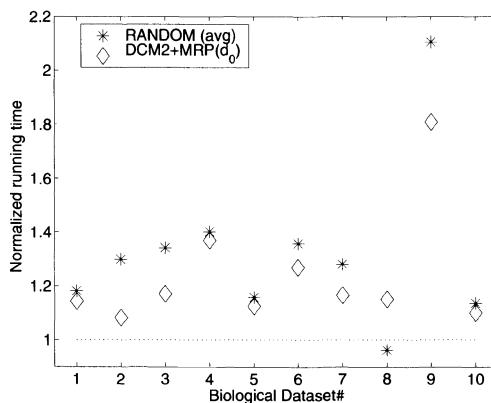


Figure 9. Comparison of running times of DCM-based methods and RANDOM (averaged over five runs) normalized with respect to the running time of DCM2 + SCM( $d_0$ ), on each of the ten biological data sets.

random decompositions, with the latter doing the worst of all. Thus, DCM2 decompositions were better than random decompositions, and SCM did a better job at assembling supertrees from such decompositions than did MRP (scores and resolution were both better). Moreover, MRP was very slow: on some data sets, the time difference was on the order of hours of computation, hours that could be used to conduct a more thorough parsimony search on the subtrees or in the OTR phase of DCM2 + SCM. (We have not run such an equal-time comparison, but we expect that the gap in parsimony scores returned by DCM2 + SCM versus the other methods would be widened.)

## 5. Results on global heuristics

Our results suggest that a DCM2 + SCM analysis is both faster and more accurate (in terms of MP scores) than the other divide-and-conquer methods studied. How then does DCM2 + SCM compare to a direct (global) heuristic approach? We expect that the DCM approach will prove better on data sets that yield good decompositions (i.e., decompose into a small number of substantially smaller data sets with good overlap), but need to ascertain how the DCM approach performs when decompositions are poor.

We selected two data sets to explore how DCM2 behaves under extreme conditions: the 500-sequence *rbcL* data set (#4) and the 816-taxon rRNA data set (#9). The first data set is a poor candidate for improvement with DCM2: it decomposes poorly and is not challenging (i.e., simple heuristic searches quickly find a solution within a few steps of the best score known). The second data set, by contrast, should enable DCM2 to yield an improvement: it decomposes well and is challenging.

We first explored various global heuristics to identify the method that performs best on the 500 *rbcL* data set; this experiment showed that the parsimony ratchet improved significantly upon other local search heuristics in PAUP\*. We therefore used the parsimony ratchet both as a base method for DCM2 + SCM (yielding a method we called the DCM2-Ratchet, which also includes a final OTR phase) and as a global optimization heuristic, and compared the performance of both methods on each data set. Finally, we explored the performance of a two-phase technique in which we used DCM2 + SCM to produce a starting tree for a subsequent search (using the ratchet), and we compared the performance of this two-phase technique to the global ratchet. For this last experiment, we compared the methods by examining their progress over the period of time needed by the global parsimony ratchet to find the best score for each of the two data sets.

### 5.1 Local improvement heuristics on the *rbcL* data set

We used the 500 *rbcL* data set to explore the performance of various local improvement heuristics as implemented in PAUP\*. These included the parsimony ratchet and heuristics of the form *fast-k max-m*. The fast-k max-m heuristic was implemented using the following commands:

```
set criterion=parsimony maxtrees=k increase=no;
hsearch start=stepwise addseq=random nreps=k swap=tbr nchuck=1
chuckscore=1;
set maxtrees=m increase=no;
hsearch start=current swap=tbr nchuck=m chuckscore=no;
```

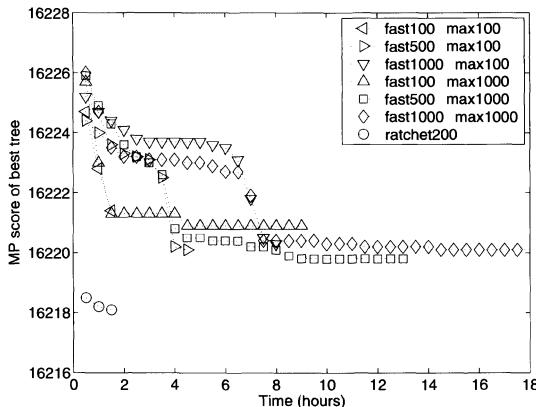


Figure 10. Comparison of heuristics on the *rbcL* data set. Each heuristic was run ten times; plotted is the average MP score at each time step.

We used 200 iterations of the ratchet (i.e., ratchet200) and varied  $k$  and  $m$  to produce the following set of heuristics for comparison:

- fast100-max100
- fast500-max100
- fast1000-max100
- fast100-max1000
- fast500-max1000
- fast1000-max1000
- ratchet200

These heuristics were studied on the 500 *rbcL* data set restricted to the parsimony-informative sites. The best score known for this data set is 16 218 steps (Nixon, 1999). We address two questions: 1) how quickly does each heuristic find the best known score and 2) how quickly does it approach this score?

We ran each heuristic ten times and recorded the average MP score at each time step. Figure 10 shows the MP score of the best tree found by each heuristic as a function of time up to the time beyond which none of the heuristics found better trees. Note that the curve for ratchet200 was always below the curves for the other methods; thus, at each time point, ratchet200 found shorter trees than all other methods. Also, ratchet200 found trees with the lowest known length within two hours, whereas the other methods could not find such short trees. Thus, the ratchet is much more effective than the fast- $k$  max- $m$  techniques in finding trees with low MP scores. Other

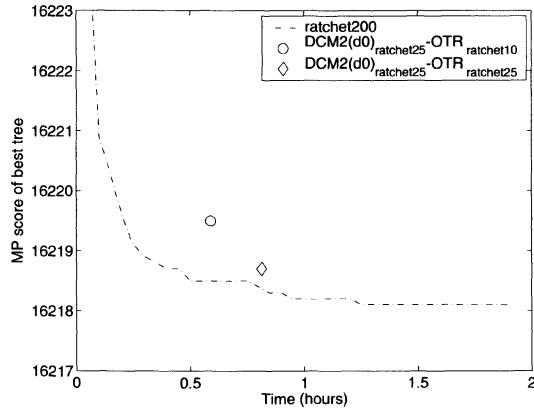


Figure 11. Comparison of DCM2-Ratchet to ratchet200 on data set #4 (averaged over ten runs). The average subproblem size is 91%.

experiments (not shown here) confirmed that the ratchet performed better on our large data sets than the fast-k max-m heuristics. We therefore used the ratchet as the base method for DCM2 and compared it against the global ratchet.

## 5.2 Global ratchet versus DCM2-Ratchet on the *rbcL* data set

We used DCM2-Ratchet with two different levels of OTR. We used the smallest possible threshold ( $d_0$ ) to obtain decompositions that minimized the maximum subproblem size. Unfortunately, even the smallest threshold yielded a huge separator of size 411, so that the two subsets in the decomposition had sizes 455 and 456 — a poor decomposition. (A larger threshold cannot help because it would produce even larger subproblems.) We compared the two OTR variants of DCM2-Ratchet against 200 iterations of the global ratchet. Each method was run ten times and the average MP score at each time step was recorded. Figure 11 shows that the DCM2-boosted variants of the ratchet were able to find trees that were almost as good as those found by the global ratchet, but not as quickly.

## 5.3 Global ratchet versus DCM2-Ratchet on the rRNA data set

For the 816-taxon rRNA data set, the DCM2 decomposition at threshold  $d_0$  produced two subproblems of sizes 369 and 450 — a good result.

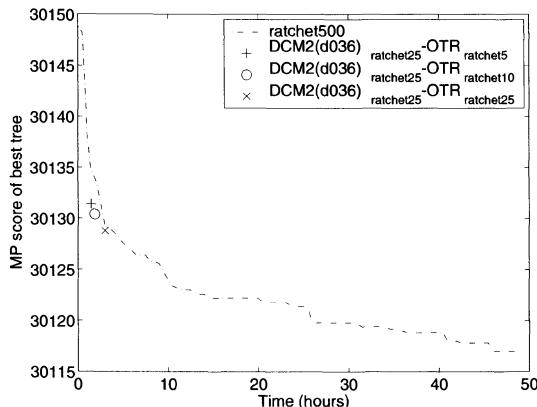


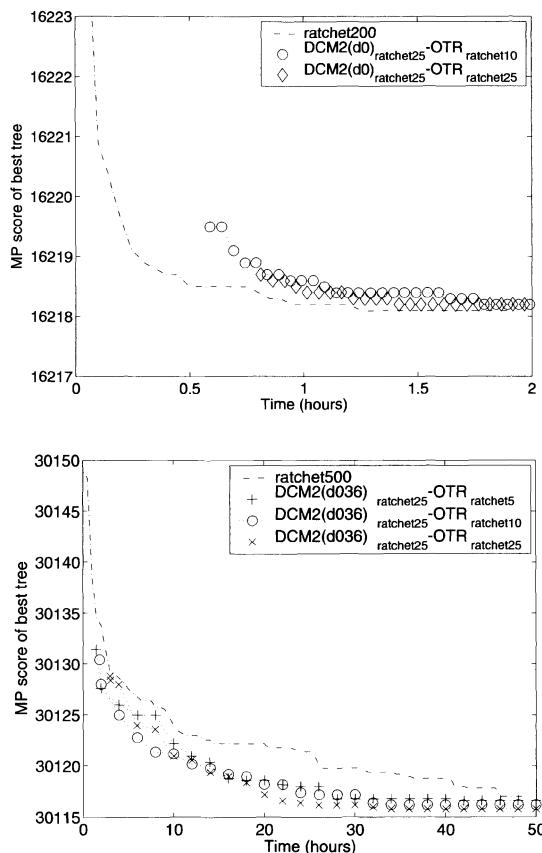
Figure 12. Comparison of DCM2-Ratchet to ratchet500 (best heuristic) on data set #9 (averaged over five runs). The average subproblem size is 36% (maximum is 60%) and the separator size is 4% of the problem size.

Unfortunately, the separator was tiny (just three nodes), which made it difficult for SCM to merge trees with sufficient accuracy. Therefore, we picked a larger separator (with 36 nodes) to get more overlap. This separator yielded three subproblems of sizes 132, 270, and 486.

Because this is a larger data set than the 500 *rbcL* one, we used 500 iterations of ratchet (ratchet500) for the global analysis. The subproblems were again computed using 25 iterations of the ratchet and three different iteration counts were used for the OTR phase: ratchet5, ratchet10, and ratchet25. Each method was run five times and the average MP score at each time step was recorded. The global ratchet finished in about 48 hours. Figure 12 shows that DCM2 found better trees than the global ratchet within the first two hours.

#### 5.4 Using DCM2-Ratchet for initialization

Because DCM2-Ratchet is fast and returns good solutions, it could prove to be useful in a two-phase optimization procedure by providing strong initial solutions from which to start a global search. To study this approach, we ran the global ratchet with the DCM2-Ratchet trees as starting trees. As might be expected, the global ratchet found better trees when started with these initial solutions than when started from scratch. Figure 13 shows the curves of (average) scores as a function of time for the global ratchet alone (started from scratch) (Figure 13a) and the global ratchet applied to the initial solutions returned by DCM2-Ratchet (Figure 13b). For instance, for the 816 rRNA data set, DCM2ratchet25-OTRratchet25 found trees with a MP score



*Figure 13.* Comparison of the global ratchet started from scratch against the same started from the DCM2-Ratchet solutions a) data set #4 (*rbcL500*) using ratchet200 and b) data set #9 (816 rDNA) using ratchet500.

of 30 115 within approximately 26 hours, whereas the best trees found by the global ratchet at the end of 48 hours had a MP score of 30 117.

## 6. Summary and conclusions

We set out to explore the potential of divide-and-conquer methods to improve the speed and accuracy of MP searches; in particular, we wanted to learn which decomposition strategies and which supertree assembly techniques work well in such approaches. Our study confirmed that divide-and-conquer methods can speed up searches for optimal MP trees, but (unsurprisingly) only when the decompositions are good.

The specifics of the divide-and-conquer strategy make a large difference. We had already shown that quartet-based methods (an extreme form of divide-and-conquer) are not competitive. We now found that random decompositions were clearly inferior to carefully crafted ones (e.g., decompositions obtained by DCM2), and that the SCM technique for merging subtrees is both more accurate and much faster than the most commonly used supertree method, MRP. Both approaches, however, usually require an OTR phase in which the initial supertree is refined into a binary tree, a phase to which more attention should be given in future work.

The significance of this study is both enhanced and limited by our use of biological data sets: we ensure relevance, but could conduct only fairly simple tests. A large simulation study is required to confirm our findings as well as to discern more subtle effects. Other research questions suggested by this study include: 1) how best to decompose data sets for which DCM2 does not produce a good decomposition, 2) how long should base methods be allowed to run on subproblems, 3) what the influence of the OTR phase on the entire process is, and 4) how can we best combine divide-and-conquer and global approaches in the style of the approach discussed in Section 5.4?

All the work in this study concerns MP, but divide-and-conquer methods (including the DCMs) are equally applicable to ML. Thus, a study of DCM-ML approaches remains to be conducted.

MP and ML scores are surrogate optimization criteria for the real goal, which is topological accuracy (unmeasurable in absence of knowledge of the “truth”). Hence, conclusions about the accuracy of reconstruction must await a simulation study in which the “true” trees are known. We conjecture that although large decreases in MP scores (or large increases in ML scores) do translate into increased topological accuracy, small changes in these scores around near-optimal values will have a nearly random effect on topological accuracy. If this is the case, there is no point in spending additional days of computation to improve a score by 0.01% and every reason to devise early termination tests.

We conclude with a caveat: in reconstructing a very large tree, such as the Tree of Life with millions of taxa, we might not have the luxury of choosing our decompositions — the data-gathering process might have made that choice for us, at least at the higher taxonomic levels. For instance, significant data might be missing for many taxa so that it is not feasible to analyze all sequences at once. In such a case, the data-set decomposition will be given to us and thus will not be adjustable (except for the breaking of large clusters). We therefore also need further large-scale evaluations of supertree methods in their traditional use.

## Acknowledgements

This work was supported by the National Science Foundation under grants ACI 00-81404 (Moret), DEB 01-20709 (Moret and Warnow), EIA 01-13095 (Moret), EIA 01-13654 (Warnow), EIA 01-21377 (Moret), EIA 01-21680 (Warnow), and EIA 99-85991 (Warnow, the SCOUT Cluster), by the David and Lucile Packard Foundation (Warnow), and by an Alfred P. Sloan Foundation Postdoctoral Fellowship in Computational Molecular Biology, U.S. Department of Energy DE-FG03-02ER63426 (Williams).

## References

- BAUM, B. R. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon* 41:3–10.
- BERRY, V., JIANG, T., KEARNEY, P., LI, M., AND WAREHAM, H. T. 1999. Quartet Cleaning: improved algorithms and simulations. In J. Nešetřil (ed.), *Algorithms — ESA'99: 7th Annual European Symposium, Prague, Czech Republic, July 1999*, Lecture Notes in Computer Science 1643:313–324. Springer-Verlag, Berlin.
- BININDA-EMONDS, O. R. P. 2003. MRP supertree construction in the consensus setting. In M. Janowitz, F.-J. Lapointe, F. R. McMorris, B. Mirkin, and F. S. Roberts (eds), *Bioconsensus*, pp. 231–242. American Mathematical Society, Providence, Rhode Island.
- BININDA-EMONDS, O. R. P., GITTLEMAN, J. L., AND PURVIS, A. 1999. Building large trees by combining phylogenetic information: a complete phylogeny of the extant Carnivora (Mammalia). *Biological Reviews* 74:143–175.
- BININDA-EMONDS, O. R. P. AND SANDERSON, M. J. 2001. An assessment of the accuracy of MRP supertree construction. *Systematic Biology* 50:565–579.
- BODLAENDER, H., FELLOWS, M., AND T. WARNOW, T. 1992. Two strikes against perfect phylogeny. In Kuich, W. (ed.), *Proceedings of the International Colloquium on Automata, Languages, and Programming ICALP'92*, Lecture Notes in Computer Science 623:273–283. Springer-Verlag, Berlin.
- BONET, M. L., STEEL, M., WARNOW, T., AND YOOSEPH, S. 1998. Better methods for solving parsimony and compatibility. *Journal of Computational Biology* 5:391–408.
- BUNEMAN, P. 1974. A characterization of rigid circuit graphs. *Discrete Mathematics* 9:205–212.
- CHASE, M. W., SOLTIS, D. E., OLMLSTEAD, R. G., MORGAN, D., LES, D. H., MISHLER, B. D., DUVALL, M. R., PRICE, R. A., HILLS, H. G., QIU, Y. L., KRON, K. A., RETTIG, J. H., CONTI, E., PALMER, J. D., MANHART, J. R., SYTSMA, K. J., MICHAELS, H. J., KRESS, W. J., KAROL, K. G., CLARK, W. D., HEDREN, M., GAUT, B. S., JANSEN, R. K., KIM, K. J., WIMPEE, C. F., SMITH, J. F., FURNIER, G. R., STRAUSS, S. H., XIANG, Q. Y., PLUNKETT, G. M., SOLTIS, P. S., SWENSEN, S. M., WILLIAMS, S. E., GADEK, P. A., QUINN, C. J., EGUILARTE, L. E., GOLENBERG, E., LEARN, G. H., GRAHAM, S. W., BARRETT, S. C. H., DAYANANDAN, S., AND ALBERT, V. A. 1993. Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene *rbcL*. *Annals of the Missouri Botanical Garden* 80:528–580.
- ERDŐS, P. L., STEEL, M. A., SZÉKELY, L. A., AND WARNOW, T. J. 1999. A few logs suffice to build (almost) all trees (Part 1). *Random Structures and Algorithms* 14:153–184.

- FELSENSTEIN, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* 17:368–376.
- FOULDS, L. R. AND GRAHAM, R. L. 1982. The Steiner problem in phylogeny is NP-complete. *Advances in Applied Mathematics* 3:43–49.
- GOLBOFF, P. 1999. Analyzing large data sets in reasonable times: solution for composite optima. *Cladistics* 15:415–428.
- GOLUMBIC, M. 1980. *Algorithmic Graph Theory and Perfect Graphs*. Academic Press, New York.
- GORDON, A. D. 1986. Consensus supertrees: the synthesis of rooted trees containing overlapping sets of labeled leaves. *Journal of Classification* 3:31–39.
- HILLIS, D. M., MORITZ, C., AND MABLE, B. 1996. *Molecular Systematics*. Sinauer Associates, Sunderland, Massachusetts.
- HUSON, D., NETTLES, S. AND WARNOW, T. 1999a. Disk-covering, a fast-converging method for phylogenetic tree reconstruction. *Journal of Computational Biology* 6:369–386.
- HUSON, D., VAWTER, L., AND WARNOW, T. 1999b. Solving large scale phylogenetic problems using DCM2. In T. Lengauer, R. Schneider, P. Bork, D. Brutlag, J. Glasgow, H.-W. Mewes, and R. Zimmer (eds), *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, pp. 118–129. AAAI Press, Menlo Park, California.
- JONES, K. E., PURVIS, A., MACLARNON, A., BININDA-EMONDS, O. R. P., AND SIMMONS, N. B. 2002. A phylogenetic supertree of the bats (Mammalia: Chiroptera). *Biological Reviews* 77:223–259.
- KIMURA, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* 16:111–120.
- LIU, F.-G. R., MIYAMOTO, M. M., FREIRE, N. P., ONG, P. Q., TENNANT, M. R., YOUNG, T. S., AND GUGEL, K. F. 2001. Molecular and morphological supertrees for eutherian (placental) mammals. *Science* 291:1786–1789.
- MAHON, A. S. 2004. A molecular supertree of the Artiodactyla. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 411–437. Kluwer Academic, Dordrecht, the Netherlands.
- MAIDAK, B. L., COLE, J. R., LILBURN, T. G., PARKER, C. P. JR, SAXMAN, P. R., FARRIS, R. J., GARRITY, G. M., OLSEN, G. J., SCHMIDT, T. M., AND TIEDJE, J. M. 2001. The RDP-II (Ribosomal Database Project). *Nucleic Acids Research* 29:173–174.
- MISHLER, B. D. 1994. Cladistic analysis of molecular and morphological data. *American Journal of Physical Anthropology* 94:143–156.
- MORET, B. M. E. 2002. Towards a discipline of experimental algorithmics. In M. H. Goldwasser, D. S. Johnson, and C. C. McGeoch (eds), *Data Structures, Near Neighbor Searches, and Methodology: Fifth and Sixth DIMACS Implementation Challenges*, pp. 197–213. American Mathematical Society, Providence, Rhode Island.
- MORET, B. M. E., ROSHAN, U., AND WARNOW, T. 2002. Sequence length requirements for phylogenetic methods. In R. Guigó and D. Gusfield (eds), *Algorithms in Bioinformatics, Second International Workshop, WABI 2002, Rome, Italy, September 17–21, 2002, Proceedings*, pp. 343–356. Springer, Berlin.
- NAKHLEH, L., MORET, B. M. E., ROSHAN, U., ST. JOHN, K., AND WARNOW, T. 2002. The accuracy of fast phylogenetic methods for large data sets. In R. B. Altman, A. K. Dunker, L. Hunter, and T. E. Klein (eds), *Pacific Symposium on Biocomputing 2002*, pp. 211–222. World Scientific Publishing Company, River Edge, New Jersey.

- NAKHLEH, L., ROSHAN, U., ST. JOHN, K., SUN, J., AND WARNOW, T. 2001. Designing fast converging phylogenetic methods. *Bioinformatics* 17:S190–S198.
- NIXON, K. C. 1999. The parsimony ratchet, a new method for rapid parsimony analysis. *Cladistics* 15:407–414.
- OLSEN, G. J., WOENSE, C. R., AND OVERBEEK, R. 1994. The winds of (evolutionary) change: breathing new life into microbiology. *Journal of Bacteriology* 176:1–6.
- POSADA, D. AND CRANDALL, K. A. 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics* 14:817–818.
- PURVIS, A. 1995. A composite estimate of primate phylogeny. *Philosophical Transactions of the Royal Society London Series B* 348:405–421.
- RAGAN, M. A. 1992. Phylogenetic inference based on matrix representation of trees. *Molecular Phylogenetics and Evolution* 1:53–58.
- RICE, K. A., DONOGHUE, M. J., AND OLMSTEAD, R. G. 1997. Analyzing large data sets: *rbcL* 500 revisited. *Systematic Biology* 46:554–563.
- SAITOU, N. AND NEI, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4:406–425.
- SOLTIS, P. S., SOLTIS, D. E., CHASE, M. W., MORT, M. E., ALBACH, D. C., ZANIS, M. J., SAVOLAINEN, V., HAHN, W. H., HOOT, S. B., FAY, M. F., AXTELL, D. C., SWENSON, S. M., PRINCE, L. M., KRESS, W. J., NIXON, K. C., AND FARRIS, J. S. 2000. Angiosperm phylogeny inferred from a combined data set of 18S rDNA, *rbcL*, and *atpB* sequences. *Botanical Journal of Linnean Society* 133:381–461.
- ST. JOHN, K., WARNOW, T., MORET, B. M. E., AND VAWTER, L. 2001. Performance study of phylogenetic methods: (unweighted) quartet methods and neighbor-joining. In S. R. Kosaraju (ed.), *Symposium on Discrete Algorithms. Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 196–205. Society for Industrial and Applied Mathematics, Philadelphia, PA.
- STEEL, M. A. 1994. The maximum likelihood point for a phylogenetic tree is not unique. *Systematic Biology* 43:560–564.
- STRIMMER, K. AND VON HAESELER, A. 1996. Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. *Molecular Biology and Evolution* 13:964–969.
- SWOFFORD, D. L. 2002. *PAUP\*. Phylogenetic Analysis Using Parsimony (\*and Other Methods). Version 4*. Sinauer, Sunderland, Massachusetts.
- TANG, J. AND MORET, B. M. E. 2003. Scaling up accurate phylogenetic reconstruction from gene-order data. *Bioinformatics* 19:i305–i312.
- WARNOW, T., B. MORET, B. M. E., AND ST. JOHN, K. 2001. Absolute convergence: true trees from short sequences. In S. R. Kosaraju (ed.), *Symposium on Discrete Algorithms. Proceedings of the Twelfth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 186–195. Society for Industrial and Applied Mathematics, Philadelphia, PA.
- WUYTS, J., VAN DE PEER, Y., WINKELMANS, T., AND WACHTER, R. D. 2002. The European database on small subunit ribosomal RNA. *Nucleic Acids Research* 30:183–185.
- YANG, Z. 1993. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution* 10:1396–1401.

## Chapter 15

### UNROOTED SUPERTREES

*Limitations, traps, and phylogenetic patchworks*

Sebastian Böcker

**Abstract:** Whereas biologists might think of rooted trees as the natural, or even the only, way to display phylogenetic relationships, this is not the case for a mathematician, to whom rooted and unrooted trees are graph-theoretical constructions that can be transformed easily into one another. An unrooted tree contains the same information as its rooted counterpart with the single exception that it does not tell you where the “evolutionary process” started. Rooting a tree is often more of an art than a science, and a pressing problem in systematic biology is precisely the exact placement of a root. In addition, many phylogenetic algorithms in fact output unrooted trees that are rooted (artificially) in a subsequent step.

From this, it is clear that finding an unrooted supertree or parent tree is of the same interest as it is for the rooted case. But, whereas a single unrooted tree can always be transformed into a rooted tree carrying the same information, this is no longer the case for collections of unrooted trees. Hence, the supertree problem for rooted trees is a special case of that for unrooted trees. As is often the case, this means that many things that can be done with rooted trees (the special case) are no longer valid for unrooted trees (the general case). In fact, the smallest possible example of a collection of unrooted trees that cannot be transformed into a collection of rooted trees is already sufficient to demonstrate that, unfortunately, many convenient features of the rooted supertree problem do not carry over to the unrooted supertree problem.

On the positive side, if the set of input trees fulfills some minimality criterion, then there exists a simple set of conditions to check whether there is exactly one parent tree for this collection. In addition, the unique parent tree, should one exist, can be constructed quickly because the set of input trees always shows a certain “patchwork” structure.

**Keywords:** parent trees; patchworks; quartet trees; unrooted supertrees;

## 1. Introduction

In the following, I will assume that all trees are unrooted unless stated otherwise. The supertree problem — that is, combining a collection of leaf-labeled trees with overlapping sets of labels (taxa) into a single “best” leaf-labeled tree — has been studied for some time (Gordon, 1986; Purvis, 1995; Sanderson *et al.*, 1998). Supertree construction methods for unrooted trees include DISK-COVERING (Huson *et al.*, 1999a, b) and dyadic closure-based approaches (Bryant and Steel, 1995; Böcker *et al.*, 2000), to name just a few. In the following, I will limit my attention to the problem of finding a “parent tree” of the input collection: given a collection  $\mathcal{F}$  of leaf-labeled trees with generally distinct, although not necessarily disjoint label sets, we want to *amalgamate* these trees into one leaf-labeled parent tree  $T$  so that all trees in  $\mathcal{F}$  are “induced” subtrees of  $T$  (see Section 2 for the distinction between supertrees and parent trees). Hence, we want to know whether such a parent tree  $T$  exists at all and, if so, whether it is determined uniquely by  $\mathcal{F}$ . Many supertree methods will try to amalgamate input trees even if no such parent tree exists. But if a unique parent tree does exist for the input collection, all reasonable supertree methods should return this parent tree.

There are several (computational) problems when trying to construct an unrooted parent tree or when determining if such a parent tree is unique:

- There exist certain limitations that apply to *all* unrooted supertree methods, and certain desirable characteristics for such supertree methods cannot be comprised in a single supertree method (see Section 3).
- The problem of determining whether there exists *at least one* parent tree of  $\mathcal{F}$  is provably difficult; that is, it is NP-complete (Steel, 1992).
- The complexity of determining whether some (known!) tree is the *unique* parent tree of a given collection is still unknown.
- The problem of determining whether some input set of trees  $\mathcal{F}$  contains an excess-free subset (see Section 5) that has a unique parent tree is also NP-complete (Böcker *et al.*, 2000).
- Finally, there exist certain collections  $\mathcal{F}$  of trees that have exponentially many parent trees in the number of trees in  $\mathcal{F}$ , as well as in the number of leaves of  $\mathcal{F}$ . Of course, this would be a trivial result if  $\mathcal{F}$  had a highly unresolved parent tree because all possible refinements of this tree would also be parent trees of  $\mathcal{F}$ . What makes the construction presented in Section 4 more compelling is that our (exponentially large) collection of parent trees will consist solely of binary trees; that is, none of these trees will have any refinement. The latter poses a threat to all supertree heuristics trying to circumvent the

runtime constraint by using a (more-or-less) intriguing construction recipe. Such heuristics might return just *one* binary parent tree  $T$  to the user that supports some hypothesis, concealing the fact that there exist exponentially many such parent trees that possibly contradict the same hypothesis.

But for certain collections of input trees, we can do better. If the set of input trees fulfills some minimality criterion, then there exists a simple set of conditions to check whether there is *exactly one* parent tree for this collection. In addition, the unique parent tree can be constructed in quadratic runtime. Here, the set of input trees always shows a certain “patchwork” structure that allows us to merge only *two* trees at a time. Although the result itself appears to be rather simple, its proof is cumbersome and lengthy. I hope that by solving the case fulfilling the minimality criterion, we can find a more efficient algorithm for a more general class of tree collections.

## 2. Definitions

Following Böcker *et al.* (2000), I begin by introducing some terminology.

- Given a tree  $T$ , a *leaf* is a vertex of  $T$  of degree one. Both a vertex that is not a leaf, as well as an edge that is not incident with a leaf are called *interior*. In the following, I assume that all interior vertices of  $T$  are of degree at least three, and that there is at least one interior vertex. Such a tree  $T$  is also called a *phylogenetic tree*.<sup>1</sup> I will use the terms “leaves”, “leaf labels”, and “taxa” interchangeably, depending on the context. If all of the interior vertices have degree three, the tree is said to be a *binary* (phylogenetic) tree.
- For a tree  $T = (V, E)$ , let  $\mathcal{L}(T) \subseteq V$  denote the set of leaf labels of  $T$ , and for a collection  $\mathcal{F}$  of such trees, let  $\mathcal{L}(\mathcal{F})$  denote the union  $\cup_{T \in \mathcal{F}} \mathcal{L}(T)$ . Recall that the number of interior edges of  $T$  never exceeds  $|\mathcal{L}(T)| - 3$ , and equality holds if and only if  $T$  is binary (see, for instance, Proposition 2.1.3 of Semple and Steel, 2003). Two trees  $T = (V, E)$  and  $T' = (V', E')$  are *isomorphic* if  $\mathcal{L}(T) = \mathcal{L}(T')$ , and there exists a bijection  $\psi : V \rightarrow V'$  that, restricted to  $\mathcal{L}(T)$ , is the identity, and that induces a bijection of  $E \rightarrow E'$ .
- Given a tree  $T$  and a subset  $L \subseteq \mathcal{L}(T)$ , I denote by  $T|_L$  the phylogenetic tree obtained from the smallest connected subgraph of  $T$  containing (the leaves labeled by)  $L$  by making this subgraph

<sup>1</sup> This definition differs from the one given in Semple and Steel (2003), but one can see easily that both definitions are in fact equivalent.

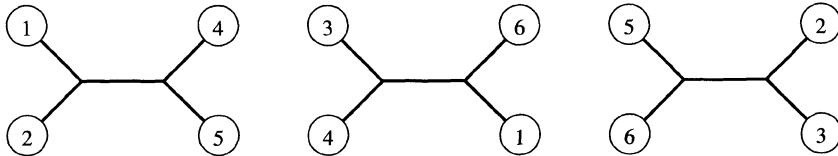


Figure 1. The three quartet trees of Example 1.

homeomorphically irreducible (i.e., by suppressing all degree-two vertices). I refer to  $T|_L$  as an *induced subtree* of  $T$  and, more specifically, as the subtree of  $T$  induced by  $L$ . Note that  $T|_L$  is binary whenever  $T$  is, and that  $(T|_L)|_L$  is (isomorphic to)  $T|_L$  for any tree  $T$  and sets  $L \subseteq L' \subseteq \mathcal{L}(T)$ .

- Given two trees  $T, T'$  with  $\mathcal{L}(T) = \mathcal{L}(T')$ , I write  $T \leq T'$  if, up to a label-preserving isomorphism,  $T$  can be obtained from  $T'$  by contracting some interior edges of  $T'$ .
- Suppose that  $\mathcal{F} := \{T_1, \dots, T_k\}$  is a collection of trees. A tree  $T$  displays  $\mathcal{F}$  if  $T_i \leq T|_{\mathcal{L}(T)}$  holds for all  $i = 1, \dots, k$ . I say also that  $T$  displays  $T'$  in case  $T$  displays  $\{T'\}$ .
- If  $T$  displays  $\mathcal{F}$  and, in addition,  $\mathcal{L}(\mathcal{F}) = \mathcal{L}(T)$  holds, then  $T$  is called a *parent tree* of  $\mathcal{F}$ .<sup>2</sup> In this case, we say also that  $\mathcal{F}$  can be *amalgamated* into the parent tree  $T$ . If this parent tree  $T$  is unique (up to isomorphisms), then I say that  $\mathcal{F}$  defines  $T$ . Note that if  $\mathcal{F}$  defines  $T$ , then  $T$  is necessarily binary.
- A *quartet tree* is a binary tree  $T$  with  $|\mathcal{L}(T)| = 4$ . I write  $xy|wz$  to denote the quartet tree that has leaves labeled  $x, y$  separated from leaves labeled  $w, z$  by its unique interior edge. For example, the quartet tree  $12|45$  is depicted on the left in Figure 1.

### 3. Limitations of unrooted supertree methods

I start this section with a simple example of a collection of trees that does not have an equivalent in the setting of rooted trees. This example is smallest possible with this property: it contains only three quartet trees on six taxa. Clearly, all input collections consisting of two trees will either have at least one leaf common to all (two) trees of the input collection, or the two trees will not share a single leaf and construction of a supertree is not reasonable. I

<sup>2</sup> Such trees are called supertrees in mathematical literature, but here I want to differentiate between the output of a supertree method and the mathematically rigid concept of “parent trees.”

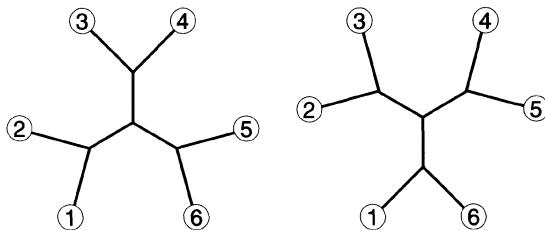


Figure 2. The parent trees  $T_+$  (left) and  $T_-$  (right) of Example 1.

will make use of Example 1 throughout this chapter to illustrate several limitations for the unrooted supertree problem.

*Example 1.* Consider the collection of quartet trees

$$(1) \quad \mathcal{Q} := \{12 | 45, 34 | 61, 56 | 23\}$$

with leaf set  $\mathcal{L}(\mathcal{Q}) = \{1, \dots, 6\}$  as displayed in Figure 1. Note that any two trees in this collection have exactly two leaves in common, and that no leaf is present in *all* trees of the collection. The key observation is that there exist (up to isomorphisms) exactly two parent trees for this collection of trees as depicted in Figure 2 (see Böcker *et al.*, 1999). I will denote the left parent tree in Figure 2 by  $T_+$ , the right parent tree by  $T_-$ .

Following the reasoning of Steel *et al.* (2000), I will now collect several properties that should be achieved (simultaneously) by all supertree methods for unrooted trees (see also Wilkinson *et al.*, 2004). Afterwards, I will show, using the above example, that no supertree method exists that can satisfy all the stated properties simultaneously.

First, it is clearly desirable that a supertree method should not rely on the ordering of the input trees. In fact, I have only talked about collections (sets) of input trees up to now that are unordered by definition. As long as we have equal confidence in all the input trees, changing the order in which we present these input trees to the supertree method should not change the output of the method.

Second, renaming the taxa should not change the output of the method. That is, if we replace a leaf label in all our input trees, then the output of the supertree method should be the old supertree, except that the new label also replaces the old label. In fact, this condition consists of two parts. First, no supertree method should force the output of a unique supertree by, say, lexicographically ordering the input taxa before applying the true supertree construction. Otherwise, renaming a taxon from “cat” to “Felidae” could

change the output of the method. Second, and more compelling, exchanging the names of two taxa in our input collection of trees should return the old supertree, except that the two taxa are also exchanged. As stated above, the supertree method cannot pre-sort the input taxa to circumvent this requirement.

Third, if the input collection has at least one parent tree, then the supertree method should return a parent tree of the collection. So, if the input trees do fit together, then the supertree method should select one of the parent trees that achieves this.

I will now show that no supertree method for unrooted trees exists that satisfies these three conditions at the same time (Proposition 1 in Steel *et al.*, 2000). Suppose our supertree method satisfies the first and second conditions above. We will see that the third condition must fail necessarily for certain input collections. To this end, consider again the collection of input trees from Example 1. Recall that this collection has exactly two parent trees (Figure 2). Suppose that our supertree method outputs the first parent tree,  $T_+$ . If we exchange taxa 2 and 6, and also taxa 3 and 5 of our input collection, the collection becomes  $\{16|43, 54|21, 32|65\}$ , which is exactly the same input collection as before because the “changed order” of elements is of no relevance. By the first condition, the method should therefore output the same parent tree  $T_+$ . But, by the second condition, the method should output the tree  $T_+$ , where leaves 2 and 6, as well as 3 and 5 are exchanged, and this is in fact the other parent tree,  $T_-$ ! The same holds true if the supertree method would output  $T_-$ . Thus, if the first *and* the second conditions hold, the output can be neither  $T_+$  nor  $T_-$ . But this means that the third condition is necessarily violated because these are the only parent trees of the input collection.

We could abandon the third condition and ask the supertree method to return *all* parent trees in case more than one exists. This might not be desirable from a phylogenetic standpoint, but it would at least allow the supertree method to return a sensible output in case there is more than one parent tree. But, besides the possible biological concerns over such an output, I will show in the next section that there might be exponentially many parent trees for certain collections of input trees. For example, for an input collection of 40 quartet trees on a set of 43 taxa, there can be as many as 1024 binary parent trees. In general, we could apply a consensus method to the set of parent trees of our input collection, but it is not clear how to carry out such calculations in a reasonable time.

Note that as soon as all the input trees have a leaf in common, we can transform the whole collection into rooted trees and use a supertree method to finally obtain an unrooted supertree by reattaching the artificial root leaf. In case a unique parent tree of such an input collection exists, this parent tree

can be found and its uniqueness can be tested quickly. But, if there is more than one choice for our pseudo-root, choosing different pseudo-roots might lead to different (rooted and unrooted) supertrees, violating the second condition we wanted our supertree method to fulfill.

#### 4. Exponentially many parent trees

Clearly, as a result of the formal definition of “parent trees” I have introduced above, the existence of (super-)exponentially many parent trees is not always a relevant problem. This is because we require a parent tree to “include all the information” of the input set of trees, but allow it to include additional information that is not supported by the input trees. As an example, define the collection of quartet trees

$$\mathcal{F}_* := \{12 | jk \mid 3 \leq j < k \leq n\}$$

for some integer  $n \geq 4$  that has leaf set  $\mathcal{L}(\mathcal{F}_*) = \{1, \dots, n\}$ . All such input trees separate leaves 1 and 2 from any two other leaves. We can see easily that  $(2n - 7)!! := 1 \cdot 3 \cdot 5 \cdots (2n - 7)$  many binary parent trees of  $\mathcal{F}_*$  exist and, therefore, also super-exponentially many parent trees in total. The binary parent trees are exactly those binary trees with leaf set  $\{1, \dots, n\}$  having an edge that separates leaves labeled 1 and 2 from leaves labeled 3, ...,  $n$ , and  $n - 4$  arbitrary other interior edges. But, there is exactly *one* parent tree of  $\mathcal{F}_*$  that is minimal with respect to  $\leq$ , and this tree (i.e., the tree with one interior edge separating leaves 1 and 2 from all other leaves) is the one every practitioner would probably be interested in. In fact, this tree is the strict consensus (see next section) of all possible parent trees of the input tree collection. Note that the production of so-called “novel clades” in supertrees has been found to be highly undesirable by several biologists (e.g., Pisani and Wilkinson, 2002; Gatesy and Springer, 2004; Wilkinson *et al.*, 2004).

I will now use Example 1 to construct an input collection of trees such that not only do exponentially many parent trees for this input tree collection exist, but also such that *every* such parent tree is necessarily binary. In addition, we will see such that these parent trees share practically no information (i.e., the strict consensus of all these parent trees is almost the star graph). This means that although the input tree collection  $\mathcal{F}$  can be amalgamated into a single parent tree, doing so would conceal the fact that there are an exponential number of parent trees that, in total, contradict the information yielded by the single parent tree completely. All heuristic supertree methods returning one, or even a few, parent trees would fall into

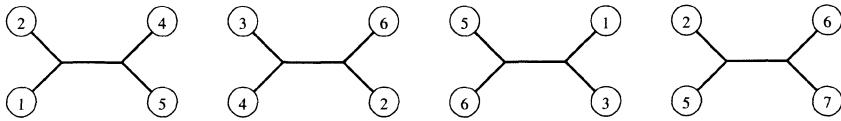
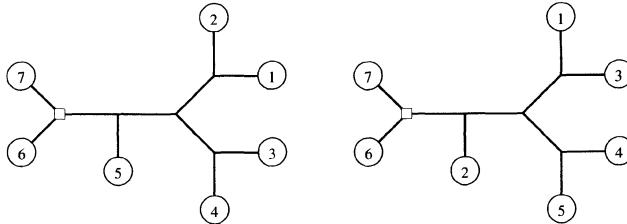


Figure 3. The four quartet trees of Example 2.

Figure 4. The parent trees  $T^+$  (left) and  $T^-$  (right); the median of the leaves 5, 6, and 7 is indicated by a square.

this trap. Let us suppose that such a heuristic returns a (randomly chosen) parent tree  $T$ , then the fact that

- the constructed parent tree is binary,
- no less refined tree obtained by contracting edges in  $T$  that is a parent tree of the input collection exists, and
- any two parent trees of  $\mathcal{F}$  differ strongly; that is, they cannot be transformed into each other by local optimization heuristics like a single branch swapping

might falsely lead the user to believe that the constructed parent tree is unique. By contrast, if a supertree method tries to construct all possible parent trees of the input collection, then it has to output a huge number of trees, and the size of the output itself makes this problem computationally hard. In other words, polynomial runtime in the input size cannot be achieved by such an approach.

*Example 2.* Consider the collection of quartet trees

$$(2) \quad \mathcal{Q}_* := \{21|45, 34|62, 56|13, 25|67\}$$

with leaf set  $\mathcal{L}(\mathcal{Q}_*) = \{1, \dots, 7\}$  as depicted in Figure 3. It is easy to check that this collection is displayed (up to isomorphism) exactly by the two trees depicted in Figure 4. In fact, I have extended Example 1 by adding a single

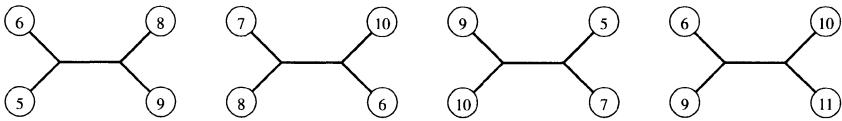


Figure 5. The four quartet trees  $\phi_4(T)$  for  $T \in \mathcal{Q}_*$ .

quartet tree  $25|67$  so that the additional leaf 7 has to be placed together with leaf 6 in both cases. Note also that I have exchanged leaves 1 and 2 to simplify the construction presented below. I will denote the left parent tree in Figure 4 by  $T^+$  and the right parent tree by  $T^-$ .

Why did I add an additional leaf and a quartet tree to Example 1? To answer this question, look at the median of the leaves 5, 6, and 7 (i.e., the unique vertex separating these three leaves) in  $T^+$ . The median of 5, 6, and 7 is the vertex adjacent to leaves 6 and 7. But this is also true for  $T^-$ ! This guarantees that if we amalgamate *any* binary tree  $T$  to  $T^+$  (or  $T^-$ ) satisfying the three conditions

1.  $5, 6, 7 \in \mathcal{L}(T)$ ,
2.  $1, 2, 3, 4 \notin \mathcal{L}(T)$ , and
3. the median of 5, 6, and 7 in  $T$  is adjacent to leaf 5,

then there exists a unique parent tree of  $T$  and  $T^+$  or of  $T$  and  $T^-$ , respectively (for a formal proof, see Böcker, 1999). And why did I swap leaves 1 and 2? Look at the median of leaves 1, 2, and 3 in  $T^+$  and  $T^-$ . In both cases, this median is adjacent to leaf 1.

We now want to “transpose” trees by adding a natural number to their leaves that are also (labeled by) natural numbers. For example, the quartet tree  $12|34$  can be transposed to the tree  $67|89$  by adding five to its leaves / labels. We can assume without loss of generality that no interior vertex is (labeled by) a natural number. Formally, given a tree  $T = (V, E)$ , I assume  $V \cap \mathbb{N} = \mathcal{L}(T)$ , where  $\mathbb{N}$  denotes the set of natural numbers. For an arbitrary tree  $T$  such that  $\mathcal{L}(T) \subseteq \mathbb{N}$ , I define the mapping  $\phi_j$  for  $j \in \mathbb{N}$  such that  $\phi_j(T)$  is the same tree as  $T$ , except that every leaf  $l$  of  $T$  is replaced by the leaf  $l + j$  in  $\phi_j(T)$ . Clearly,  $\phi_0(T)$  is (isomorphic to) the input tree  $T$ . As an example, Figure 5 depicts the four trees  $\phi_4(T)$  for  $T \in \mathcal{Q}_*$ . Note that both  $\phi_4(T^+)$  and  $\phi_4(T^-)$  satisfy the conditions stated in the previous paragraph: in both cases, the median of 5, 6, and 7 is adjacent to leaf 5.

Define the transposed collections of quartet trees

$$\mathcal{Q}_k := \{\phi_{4k}(T) \mid T \in \mathcal{Q}_*\}$$

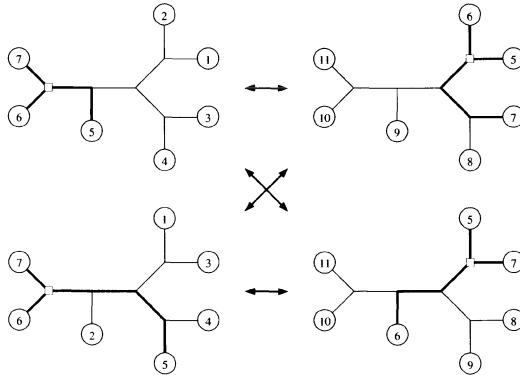


Figure 6. The four trees  $T_k^i$  for  $k = 0, 1$  (left, right) and  $i \in \{+, -\}$  (top, bottom); the median of the leaves 5, 6, and 7 is indicated by a square.

as well as the trees  $T_k^+ := \phi_{4k}(T^+)$  and  $T_k^- := \phi_{4k}(T^-)$ . Clearly,  $T_k^+$  and  $T_k^-$  are the unique parent trees of the collection  $\mathcal{Q}_k$ . As an example, the trees  $T_0^+$ ,  $T_0^-$ ,  $T_1^+$ , and  $T_1^-$  are depicted in Figure 6.

Now consider the collection of quartet trees  $\mathcal{Q}_0 \cup \mathcal{Q}_1$  with leaf set  $\mathcal{L}(\mathcal{Q}_0 \cup \mathcal{Q}_1) = \{1, \dots, 11\}$ . One can show that any parent tree of this collection is the amalgamation of a parent tree  $T_0^+$  or  $T_0^-$  of  $\mathcal{Q}_0$ , and a parent tree  $T_1^+$  or  $T_1^-$  of  $\mathcal{Q}_1$ . These four parent trees are depicted in Figure 7. This example indicates how the trees  $T_j^+$  and  $T_j^-$  can be used as “binary switches” to construct exponentially many parent trees. Eight parent trees exist for the collection  $\mathcal{Q}_0 \cup \mathcal{Q}_1 \cup \mathcal{Q}_2$ , and these parent trees can be constructed by amalgamating any parent tree of  $\mathcal{Q}_0 \cup \mathcal{Q}_1$  in Figure 7 with either  $T_2^+$  or  $T_2^-$ . By repeating this process, we see that the collection

$$(3) \quad \mathcal{Q}_k^* := \mathcal{Q}_0 \cup \mathcal{Q}_1 \cup \dots \cup \mathcal{Q}_k$$

with  $4k + 4$  elements and leaf set  $\{1, \dots, 4k + 7\}$  has exactly  $2^{k+1}$  parent trees, of which every one is binary (for a formal proof, see Böcker, 2002).

Finally, we want to know what information is shared by all parent trees of the collection  $\mathcal{Q}_k^*$ . The *strict consensus* of a collection  $\mathcal{F}$  of trees is a tree  $T^*$  such that  $T^* \leq T$  holds for all  $T \in \mathcal{F}$ , and  $T^*$  is maximal with respect to this condition. Now, what is the strict consensus tree  $T^*$  of  $\mathcal{Q}_k^*$ ? This is in fact a unique tree:

**Lemma 1.** *For  $\mathcal{Q}_k^*$  as defined in Equation 3, let  $\mathcal{F}_k$  denote the set of parent trees of  $\mathcal{Q}_k^*$ . The strict consensus of all parent trees in  $\mathcal{F}_k$  is the phylogenetic*

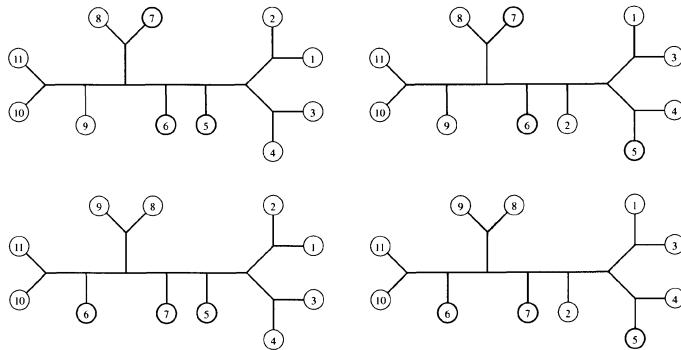


Figure 7. The four parent trees of  $\mathcal{Q}_0 \cup \mathcal{Q}_1$ .

tree  $T^*_k$  with leaf set  $\mathcal{L}(T^*_k) = \{1, \dots, 4k + 7\}$  having a single interior edge separating leaves  $\{1, \dots, 4k + 5\}$  from leaves  $\{4k + 6, 4k + 7\}$ .

I will omit the proof of this lemma, and just note that induction on  $k$  can be used to prove the claim. Furthermore, all consensus methods applied to the set of parent trees  $\mathcal{F}_k$  face the following problem: every interior edge of the potential consensus tree separates the leaves into two sets, both of cardinality larger or equal two. Except for the edge separating  $4k + 6, 4k + 7$  from all other leaves, all such interior edges are supported either by none of the input parent trees and should never be included in the output consensus tree, or by exactly half of the input parent trees and contradicted by the other half! Thus, the only reasonable output for any consensus method seems to be the strict consensus tree described above — otherwise, the output would be completely arbitrary. For example, if we choose one of the input parent trees randomly as the output of the consensus tree method, every edge of the output tree would be supported by half of the input trees, and this is as good as it gets.

Note that the input collections  $\mathcal{Q}^*_k = \mathcal{Q}_0 \cup \dots \cup \mathcal{Q}_k$  have no equivalent in the setting of rooted trees. The leaf sets of the subcollections  $\mathcal{Q}_j$  are “walking” in the sense that  $\mathcal{L}(\mathcal{Q}_j) \cap \mathcal{L}(\mathcal{Q}_{j+1}) = \{4j + 5, 4j + 6, 4j + 7\}$  contains exactly three elements, whereas all other intersections of leaf sets  $\mathcal{L}(\mathcal{Q}_j) \cap \mathcal{L}(\mathcal{Q}_k)$  are empty for  $|j - k| > 1$ . The presented construction cannot work when all leaf sets have at least one element in common.

It is worth mentioning that the constructed collections of quartet trees  $\mathcal{Q}^*_k$  are excess-free (see the next section). Thus,  $\mathcal{Q}^*_k$  is of minimal cardinality in the sense that all collections of quartet trees of smaller cardinality, but with same leaf set, must have at least one non-binary parent tree.

## 5. Solution to the parent tree problem for the minimum case

As I mentioned in Section 1, the general problem of finding a parent tree is NP-complete, and the complexity of deciding whether some given tree is the unique parent tree of a collection is unknown. In this section, I finally present a positive result: in those cases where some minimality criterion is satisfied, it is possible to construct a unique parent tree of an input collection of unrooted trees in polynomial time, even when there is no single leaf shared by all trees of the input collection. If such a unique parent tree exists, it is also the “most natural” output of any supertree method.

This section reviews work found in more detail in Böcker (1999), and all theorems and lemmata are drawn from that work unless stated otherwise.

Let  $\mathcal{F} = \{T_1, T_2, \dots, T_k\}$  denote a collection of binary trees. If  $T$  is the unique parent tree of the collection  $\mathcal{F}$ , then we can show that

$$(4) \quad |\mathcal{L}(T)| - 3 \leq \sum_{j=1}^k (|\mathcal{L}(T_j)| - 3)$$

must always hold. I refrain from giving a formal proof here (see, for example, Böcker *et al.*, 1999), but will explain instead the idea behind this formula. Every binary tree  $T$  has exactly  $|\mathcal{L}(T)| - 3$  interior edges. Thus, Equation 4 compares the number of interior edges of the parent tree with the sum of interior edges of the input trees. Now, for every interior edge of the parent tree, at least one interior edge in at least one input tree should exist that “distinguishes” the interior edge of the parent tree. Otherwise, we could remove the interior edge from the parent tree, and the resulting tree would still be a parent tree, violating our assumption of uniqueness.

Equation 4 suggests that particular attention should be paid to the *minimum case*; that is, the case where equality holds. To this end, I define the excess of the collection  $\mathcal{F}$  by

$$\text{exc}(\mathcal{F}) := |\mathcal{L}(\mathcal{F})| - \sum_{T \in \mathcal{F}} (|\mathcal{L}(T)| - 3) - 3,$$

and I say that  $\mathcal{F}$  is *excess-free* if  $\text{exc}(\mathcal{F}) = 0$  holds. Clearly,  $\text{exc}(\{T\}) = |\mathcal{L}(T)| - (|\mathcal{L}(T)| - 3) - 3 = 0$  holds for every binary tree  $T$ . The main result of this section now follows.

**Theorem 1.** *Suppose  $\mathcal{F}$  is a non-empty and excess-free collection of binary trees. Then  $\mathcal{F}$  uniquely defines a parent tree  $T$  if and only if  $\#\mathcal{F} = 1$  or there*

exists a bipartition of  $\mathcal{F}$  into two proper, disjoint subsets  $\mathcal{F}_1, \mathcal{F}_2 \subset \mathcal{F}$  such that  $\mathcal{F}_j$  is excess-free and uniquely defines a parent tree  $T_j$  for  $j = 1, 2$ , and  $T$  is the unique parent tree of  $T_1, T_2$ .

The non-trivial part of this theorem is equivalent to the following lemma.

**Lemma 2.** *Given a collection of binary trees  $\mathcal{F}$  that is excess-free, uniquely defines a parent tree, and has cardinality at least two, then there exist two distinct trees  $T, T' \in \mathcal{F}$  such that the collection  $\{T, T'\}$  is excess-free and uniquely defines a parent tree.*

Check carefully what Theorem 1 says — and even more importantly — what it *does not* say. First, the theorem guarantees the existence of a *unique* parent tree. But if the conditions of the theorem are violated, there might be either no parent tree at all or more than one parent tree. Using Theorem 1, we cannot distinguish between these two cases! But this does not come as a surprise because, as I stated above, the problem of deciding if there exists at least one parent tree of a collection of input trees is NP-complete, and below I present an algorithm with polynomial runtime to check whether the conditions of the theorem are fulfilled. Second, Theorem 1 deals with *binary* input trees only, and cannot be extended to non-binary input trees. In the following, I assume that our input collection consists solely of binary trees. Third, the theorem deals with (excess-free) tree collections of *minimum size*, and cannot be extended to minimal tree collections; that is, collections that define a unique parent tree, while any subcollection on the same leaf set has at least two parent trees (see Example 3 below). Finally, it should be obvious that biological data will almost always violate the strict conditions of the theorem, prohibiting its direct application. The idea is that, by using Theorem 1, we might be able to prove that certain supertree methods have certain desirable properties.

*Example 3 (Steel, 1992).* The set of quartet trees  $\mathcal{Q} := \{12|35, 24|57, 13|47, 34|56, 15|67\}$  has a unique parent tree with leaf set  $\{1, \dots, 7\}$ ; namely, the caterpillar depicted at the top of Figure 8. For every subcollection of  $\mathcal{Q}$  of cardinality two to four, at least two parent trees exist.

Actually, Steel (1992) showed only that no subcollections of cardinality four that have a unique parent tree exist, which is sufficient to show that  $\mathcal{Q}$  is minimal as defined above. But, we can see easily that all *but one* subcollections  $\mathcal{Q}' \subseteq \mathcal{Q}$  of cardinality  $|\mathcal{Q}'| \in \{2, 3\}$  have positive excess and, hence, cannot have a unique parent tree; whereas  $\mathcal{Q}' := \{12|35, 24|57, 13|47\}$ , being the unique excess-free subcollection, also has two parent

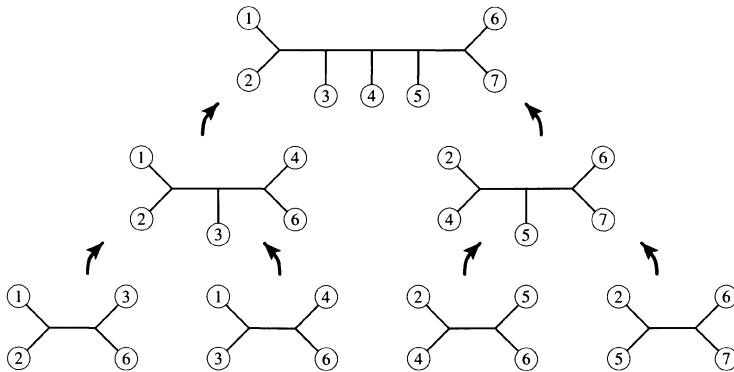


Figure 8. Hierarchy of amalgamation for the input collection  $\mathcal{F} := \{12|36, 13|46, 24|56, 25|67\}$  from Example 4.

trees. In particular, this means that no two distinct trees  $T, T' \in \mathcal{F}$  exist such that the collection  $\{T, T'\}$  defines a parent tree (compare to Lemma 2).

Lemma 2 indicates how we can reconstruct a parent tree for the minimum case. Given our excess-free input collection  $\mathcal{F}$ , we search for two trees  $T, T' \in \mathcal{F}$  such that the collection  $\{T, T'\}$  is excess-free and uniquely defines a parent tree. Then, we amalgamate  $T$  and  $T'$  into a parent tree  $T''$  (in fact, Lemma 4 below allows us to check whether a unique parent tree exists without actually constructing it), and replace  $T$  and  $T'$  in  $\mathcal{F}$  by  $T''$ . In so doing, we have reduced the cardinality of the set  $\mathcal{F}$  by one, and we repeat the process until only one parent tree is left. Surprisingly, this means that we construct a rooted tree (or, equivalently, a hierarchy) with leaves labeled by the trees of our input collection that tells us the order in which the input trees must be amalgamated to construct the parent tree of our collection (see Example 4). Böcker *et al.* (2000), present an algorithm with runtime  $O(|\mathcal{L}(\mathcal{F})|^2)$  to reconstruct the unique parent tree of an excess-free input collection.

*Example 4.* Let  $\mathcal{F} := \{12|36, 13|46, 24|56, 25|67\}$  denote a collection of quartet trees. One can show that the unique parent tree of this collection is the caterpillar tree depicted at the top of Figure 8. In addition, Figure 8 displays an amalgamation hierarchy that shows how the input trees can be amalgamated pairwise.

But why does this simple algorithm work? That is, why can we amalgamate *any* two trees of the input collection that form an excess-free subcollection? Is there no possibility that we will run into a dead end by

amalgamating two “wrong” trees in the beginning such that we end up with a partially merged set of trees and can no longer find two trees to merge?

In fact, Lemma 2 is sufficient to prove that the algorithm will return the unique parent 0 in case it exists. To this end, note that the parent tree  $T^*$  of  $\mathcal{F}$  necessarily displays  $T''$ , where  $T''$  denotes the unique parent tree of  $\{T, T'\}$ . By contrast, all trees that display  $T''$  must also display  $T$  and  $T'$ . This implies that the collection  $\mathcal{F}' := \mathcal{F} - \{T, T'\} \cup \{T''\}$  also has the unique parent tree  $T^*$ . Finally, we can show that the collection  $\mathcal{F}'$  is also excess-free, and the lemma guarantees that we can find two trees in  $\mathcal{F}'$  that we can amalgamate, and so on.

But a more exhaustive answer to these questions lies in a certain structure that subcollections of our input collection exhibit. We need a new mathematical tool to capture the concept behind this structure. Let  $X$  denote an arbitrary set, and let  $\mathcal{P}(X)$  denote the set of all subsets of  $X$ . I say that  $C \subseteq \mathcal{P}(X)$  (that is,  $C \subseteq X$  holds for all  $C \in C$ ) is a *patchwork* if the following condition is satisfied:

$$\text{If } A, B \in C \text{ and } A \cap B \neq \emptyset, \text{ then } A \cap B \in C \text{ and } A \cup B \in C.$$

*Example 5.* Let  $X := \mathbf{R}$  denote the set of real numbers. Then, the set of all intervals in  $X$ ,  $C := \{[a, b] \mid a, b \in \mathbf{R}\}$  forms a patchwork. Given two intervals  $[a, b]$  and  $[c, d]$ , these intervals are either disjoint, or  $[a, b] \cap [c, d]$  and  $[a, b] \cup [c, d]$  both form intervals. The same holds true for open and half-open intervals, and if  $X$  is the set of rational or natural numbers, or the set of integers.

The following two lemmata show that such patchwork structures appear naturally in the context of constructing parent trees from minimum collections, and that the elements of this patchwork are of interest when trying to reconstruct the parent tree.

**Lemma 3 (Lemma 3.10 of Böcker *et al.*, 1999).** *Given an excess-free collection  $\mathcal{F}$  of binary trees with a unique parent tree, the subcollections of  $\mathcal{F}$  that are excess-free form a patchwork.*

**Lemma 4.** *Given an excess-free collection  $\mathcal{F}$  of binary trees with a unique parent tree, a subcollection  $\mathcal{F}' \subseteq \mathcal{F}$  is excess-free if and only if the collection  $\mathcal{F}'$  has a unique parent tree.*

The latter lemma tells us that to check whether some subcollection of our input collection  $\mathcal{F}$  (that has a unique parent tree) also defines some unique parent tree, it is sufficient to calculate the excess of the collection. For

Example 4, we know already that subcollections  $\{12|36, 13|46\}$  and  $\{24|56, 25|67\}$  are excess-free. In addition, the collection  $\{12|36, 13|46, 24|56\}$  is excess-free and, hence, has a unique parent tree: the caterpillar on leaves 1, ..., 6. This allows for an alternative hierarchy to reconstruct the unique parent tree of the collection. Also note that if we replace the input tree  $24|56$  by  $24|57$  in Example 4, then the “hierarchy of amalgamation” presented in Figure 8 displays the unique way of constructing the parent tree.

Patchworks were introduced in Böcker and Dress (2001), and several equivalent conditions were introduced for a patchwork to be ample. I call  $C \subseteq \mathcal{P}(X)$  *ample* if the following condition holds:

$$\text{If } A, C \in \mathcal{C} \text{ satisfies } A \subset C, \text{ and there exists no } B \in \mathcal{C} \text{ with } A \subset B \subset C, \text{ then } C \setminus A \in \mathcal{C}.$$

Recall that “ $A \subset B$ ” denotes a proper subset  $A \subseteq B$  with  $A \neq B$ . A set  $C \subseteq \mathcal{P}(X)$  is called a *hierarchy* if it satisfies:

$$\text{If } A, B \in C, \text{ then } A \subseteq B, B \subseteq A, \text{ or } A \cap B = \emptyset \text{ holds.}$$

A hierarchy  $C \subseteq \mathcal{P}(X)$  is called *maximal* if there exists no hierarchy  $C' \subseteq \mathcal{P}(X)$  such that  $C \subset C'$ . Recall that there is a one-to-one correspondence between hierarchies  $C \subseteq \mathcal{P}(X)$  and rooted trees with leaf set  $X$ .

**Theorem 2 (Theorem 1 of Böcker and Dress, 2001).** *A patchwork  $C \subseteq \mathcal{P}(X)$  contains a maximal hierarchy if and only if  $C$  is ample;  $\emptyset, X \in C$ ; and  $\{x\} \in C$  for all  $x \in X$  holds.*

In view of  $\text{exc}(\{T\}) = 0$  and  $\text{exc}(\mathcal{F}) = 0$  for our input collection  $\mathcal{F}$ , this implies that the excess-free subcollections of  $\mathcal{F}$ ,

$$C(\mathcal{F}) := \{\mathcal{F}' \subseteq \mathcal{F} \mid \text{exc}(\mathcal{F}') = 0\}$$

form an ample patchwork if and only if  $C(\mathcal{F}) \cup \{\emptyset\}$  contains a maximal hierarchy.

We can use the theory of patchworks to prove some non-trivial equivalences. Theorem 3 of Böcker and Dress (2001) states more equivalent conditions for a patchwork to be ample. I utilized these conditions in Böcker (1999) to show that Theorem 1 and Theorem 3 below are in fact equivalent, and that the non-trivial part of these theorems is equivalent to Lemma 2.

Proving the results presented up to this point is possible almost completely using combinatorics on leaf sets without referring explicitly to the parent tree  $T$  of the input collection. Unfortunately, this is not enough to

prove Theorem 1. For its proof as presented in Böcker (1999), many more mathematical tools would have to be introduced. But even then, a lengthy “proof residual” remains, going beyond the scope of this chapter. Thus, I will present only some of the concepts and ideas used for proving Theorem 1 because they are of interest even without the proof itself.

In the following, we want to take the structure of the parent tree into account. To this end, we suppose that we know the parent tree in advance. From that, we can derive certain necessary conditions on our input collection, and, if an input collection violates any of these conditions, we know that our assumption of a unique parent tree must be violated as well.

Limiting ourselves to quartet trees only can reduce the complexity of the formalism used. However, because every binary tree  $T$  can be encoded uniquely using  $|\mathcal{L}(T)| - 3$  quartet trees, this does not limit the results obtained. For example, the caterpillar from Example 4 can be encoded using the collection  $\mathcal{F}$  of four elements provided in the example. To this end, let us suppose from now on that we are given a collection of quartet trees  $\mathcal{Q}$ .

The next complexity reduction comes from the following. I say that an (interior) edge  $e$  of a tree  $T$  *displays* a quartet tree  $ab|cd$  with  $a, b, c, d \in \mathcal{L}(T)$  if, by removing the edge  $e$  from  $T$ , the resulting graph contains  $a, b$  in one connected component and  $b, c$  in the other component. I say that the quartet tree  $ab|cd$  *distinguishes*  $e$  if  $e$  is the unique edge of  $T$  that displays  $ab|cd$ . Now suppose that  $T$  is the unique parent tree of the collection  $\mathcal{Q}$  of quartet trees. One can show easily that every interior edge  $e$  of  $T$  is distinguished by at least one quartet tree in  $\mathcal{Q}$ . In addition, every quartet tree distinguishes at most one interior edge of  $T$ .

If we now assume that our input collection is excess-free, then there are exactly as many interior edges in the binary parent tree  $T$  as there are quartet trees in  $\mathcal{Q}$ . Thus, every quartet tree in  $\mathcal{Q}$  distinguishes *exactly one* interior edge of  $T$ . This means that we can assume in the following that, given an excess-free collection of quartet trees  $\mathcal{Q}$  and parent tree  $T$ , we can construct a one-to-one mapping  $q$  from the interior edges of  $T$  (denoted  $E^*$ ) onto the quartet trees  $\mathcal{Q}$ . In addition, we can assume that  $q(e)$  distinguishes  $e$  for all  $e \in E^*$ ; such mappings  $q$  are called *tight* in Böcker (1999) and Böcker *et al.* (1999). If no such mapping exists, the parent tree is not unique, and, as noted above, constructing all parent trees of  $\mathcal{Q}$  might be computationally hard.

I now formulate Theorem 1 in a way that allows us to use the tools introduced above:

**Theorem 3.** *Given a quartet encoding  $q$  of a binary tree  $T$  with interior edges  $E^*$ , the tree  $T$  is the unique parent tree of the collection  $\mathcal{Q} := q(E^*)$  if and only if (a)  $q$  is tight and (b) the patchwork  $C$  of subsets  $F \subseteq E^*$  satisfying  $\text{exc}(q(F)) = 0$  is ample.*

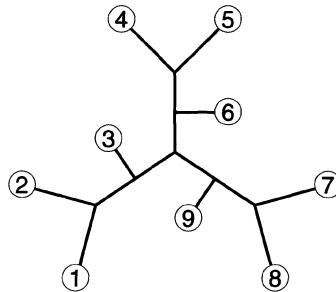


Figure 9. Parent tree from Example 6.

By looking at interior edges of  $T$  instead of quartet trees in  $\mathcal{Q}$ , we have gained the possibility to derive certain features of sets of interior edges. In fact, patchwork structures appear for a second time. Given a tree  $T$  with interior edges  $E^*$ , I say that  $F \subseteq E^*$  is a *patch* if the subgraph induced by  $F$  (that is, removing vertices  $v$  if no edge  $e \in F$  containing  $v$  exists) is connected and, therefore, a tree. Now, one can see easily that the set of all patches of a given tree  $T$  forms a patchwork too! One useful application of this concept is that excess-free subsets form patches in the tree domain:

**Lemma 5 (Lemma 3.6 iii of Böcker et al., 1999).** *Suppose  $q$  is a tight quartet encoding of a tree  $T$  and that  $F \subseteq E^*$  is a subset of interior edges. If the collection of quartet trees  $q(F)$  is excess-free, then  $F$  is a patch.*

A simple application of this lemma is the following example.

*Example 6.* Let  $q$  denote a tight quartet encoding of the tree  $T$  depicted in Figure 9. Let  $F_1, F_2$  denote a partitioning of  $E^*$ ; that is, subsets  $F_1, F_2 \subseteq E^*$  of interior edges of  $T$  with  $F_1 \cap F_2 = \emptyset$  and  $F_1 \cup F_2 = E^*$ . Clearly, there exist no such sets  $F_1, F_2$  with  $|F_1| = |F_2| = 3$ , and  $F_1$  and  $F_2$  are both patches. This implies that for all collections of quartet trees  $\mathcal{Q}$  that have the unique parent tree  $T$  depicted in Figure 9, there is no partitioning of  $\mathcal{Q}$  into two subsets  $\mathcal{Q}_1, \mathcal{Q}_2$  of cardinality three that are both excess-free. In turn, this implies that no two trees  $T_1, T_2$  with  $|L(T_1)| = |L(T_2)| = 6$  exist such that  $T$  is the unique parent tree of the collection  $\{T_1, T_2\}$ .

As mentioned above, applying our results to biological data will not lead to satisfactory results because the strict conditions of Theorem 1 will be violated almost always. But the theorem can be used to prove that the *dyadic closure* (Colonius and Schulze, 1981; Dekker, 1986) can be guaranteed to

return the “correct” answer in certain cases. Suppose  $\mathcal{Q}$  is an arbitrary collection of quartet trees, and that all these quartet trees are induced subtrees of some (unknown) parent tree  $T$ . The dyadic closure applies two simple rules to any two trees in the collection  $\mathcal{Q}$  to infer other quartet trees that are also induced subtrees of  $T$ , and adds these trees to  $\mathcal{Q}$ . The dyadic closure of any collection  $\mathcal{Q}$  can be computed in  $O(n^5)$  time for  $n := |\mathcal{L}(\mathcal{Q})|$  (see Erdős *et al.*, 1999). In Böcker *et al.* (2000), the dyadic closure operation was combined with the Berry-Gascuel construction (Berry and Gascuel, 1997) to form the DYADIC TREE CONSTRUCTION algorithm.

The “performance guarantee” mentioned above is as follows. Suppose we are given a collection of quartet trees  $\mathcal{Q}$ . If  $\mathcal{Q}$  contains an (unknown) subset  $\mathcal{Q}' \subseteq \mathcal{Q}$  with leaf set  $\mathcal{L}(\mathcal{Q}') = \mathcal{L}(\mathcal{Q})$  that is excess-free and has a unique parent tree  $T$ , then the DYADIC TREE CONSTRUCTION algorithm will (in polynomial runtime) either construct  $T$  or output that no parent tree of  $\mathcal{Q}$  exists. But in case no such subset exists, the algorithm might be able to construct one or more parent trees, to decide that no such parent tree can exist, or get stuck without providing such information. It would be nice to decide upfront if a collection of quartet trees  $\mathcal{Q}$  contains a subset  $\mathcal{Q}' \subseteq \mathcal{Q}$  that is excess-free and has a unique parent tree  $T$ , but, unfortunately, this problem is also NP-complete (Böcker *et al.*, 2000).

## 6. Conclusions

The problem of constructing unrooted supertrees or parent trees comprises certain risks that have no equivalent in the rooted tree setting. In particular, we have seen that unrooted supertree methods cannot achieve certain desirable properties simultaneously, and that there can be a large number of supertrees containing contradicting information. These problems can be circumvented by choosing input collections such that all input trees share one or more leaves. Finally, we have seen how to derive performance guarantees for an intuitively stringent supertree method in Section 5. Although Theorem 1 cannot be applied to biological data, it might allow for other performance guarantees of this type, or for loosening the “suggestion” that all input trees should share a leaf.

Is the problem of exponentially many parent trees, as introduced in Section 4, likely to arise in practice? This depends strongly on the kind as well as the “amount” of input data provided to the given supertree method. If an unrooted supertree method tries to reconstruct a parent tree given a collection of trees of *minimal cardinality* then, as a result of the small size of Example 2, it is possible that one or more subcollections of trees analogous to  $\mathcal{Q}^*$  (from Equation 3) can be found in the input collection. But if the input

collection is not of minimal size, it is unclear if and how frequently these problems might arise. Yet, the existence of this phenomenon suggests the possibility of circumventing it in the first place, for example by choosing input trees with at least one leaf in common.

## Acknowledgements

I want to thank Mike Steel who suggested the outline of this chapter and gave helpful feedback on the manuscript, and Olaf Bininda-Emonds for not losing hope when I was late with my submission. I also thank the latter and my two anonymous reviewers for many helpful suggestions that improved the readability of this chapter. Support was provided by the Deutsche Forschungsgemeinschaft (BO 1910/1-1) within the Computer Science Action Program.

## References

- BERRY, V. AND GASCUEL, O. 1997. Inferring evolutionary trees with strong combinatorial evidence. In T. Jiang and D. T. Lee (eds), *Computing and Combinatorics: Third Annual International Conference, COCOON '97, Shanghai, China, August 20–22, 1997: Proceedings*. Lecture Notes in Computer Science 1276:111–123. Springer, Berlin.
- BÖCKER, S. 1999. *From Subtrees to Supertrees*. Ph.D. thesis, Universität Bielefeld, Germany. (Available from <http://archiv.ub.uni-bielefeld.de/disshabi/2000/0001.ps>)
- BÖCKER, S. 2002. Exponentially many supertrees. *Applied Mathematical Letters* 15:861–865.
- BÖCKER, S., BRYANT, D., DRESS, A. W., AND STEEL, M. A. 2000. Algorithmic aspects of tree amalgamation. *Journal of Algorithms* 37:522–537.
- BÖCKER, S. AND DRESS, A. W. 2001. Patchworks. *Advances in Mathematics* 157:1–21.
- BÖCKER, S., DRESS, A. W., AND STEEL, M. A. 1999. Patching up  $X$ -trees. *Annals of Combinatorics* 3:1–12.
- BRYANT, D. AND STEEL, M. A. 1995. Extension operations on sets of leaf-labelled trees. *Advances in Applied Mathematics* 16:425–453.
- COLONIUS, H. AND SCHULZE, H.-H. 1981. Tree structures for proximity data. *British Journal of Mathematical and Statistical Psychology* 34:167–180.
- DEKKER, M. 1986. *Reconstruction Methods for Derivation Trees*. Master's thesis, Vrije Universiteit, Amsterdam, the Netherlands.
- ERDŐS, P. L., STEEL, M. A., SZÉKELY, L. A., AND WARNOW, T. J. 1999. A few logs suffice to build (almost) all trees (Part 1). *Random Structures and Algorithms* 14:153–184.
- GATESY, J. AND SPRINGER, M. S. 2004. A critique of matrix representation with parsimony supertrees. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 369–388. Kluwer Academic, Dordrecht, the Netherlands.
- GORDON, A. 1986. Consensus supertrees: The synthesis of rooted trees containing overlapping sets of labelled leaves. *Journal of Classification* 3:335–348.
- HUSON, D., NETTLES, S. AND WARNOW, T. 1999a. Disk-covering, a fast-converging method for phylogenetic tree reconstruction. *Journal of Computational Biology* 6:369–386.

- HUSON, D., VAWTER, L., AND WARNOW, T. 1999b. Solving large scale phylogenetic problems using DCM2. In T. Lengauer, R. Schneider, P. Bork, D. Brutlag, J. Glasgow, H.-W. Mewes, and R. Zimmer (eds), *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, pp. 118–129. AAAI Press, Menlo Park, California.
- PISANI, D. AND WILKINSON, M. 2002. Matrix representation with parsimony, taxonomic congruence, and total evidence. *Systematic Biology* 51:151–155.
- PURVIS, A. 1995. A composite estimate of primate phylogeny. *Philosophical Transactions of the Royal Society of London, Series B* 348:405–421.
- SANDERSON, M. J., PURVIS, A., AND HENZE, C. 1998. Phylogenetic supertrees: assembling the trees of life. *Trends in Ecology and Evolution* 13:105–109.
- SEMPLE, C. AND STEEL, M. 2003. *Phylogenetics*. Oxford University Press, Oxford.
- STEEL, M. A. 1992. The complexity of reconstructing trees from qualitative characters and subtrees. *Journal of Classification* 9:91–116.
- STEEL, M. A., DRESS, A. W., AND BÖCKER, S. 2000. Simple but fundamental limitations on supertree and consensus tree methods. *Systematic Biology* 49:363–368.
- WILKINSON, M., THORLEY, J. L., PISANI, D., LAPOINTE, F.J., AND MCINERNEY, J. O. 2004. Some desiderata for liberal supertrees. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 227–246. Kluwer Academic, Dordrecht, the Netherlands.

## Chapter 16

# THE CLADISTICS OF MATRIX REPRESENTATION WITH PARSIMONY ANALYSIS

Harold N. Bryant

**Abstract:** The construction of supertrees using matrix representation with parsimony (MRP) is equivalent operationally to the construction of cladograms using cladistic analysis of character data. However, the validity of MRP as a phylogenetic method has been questioned because the data used to construct MRP supertrees are the topologies of trees rather than character data. The consistency of MRP analysis with the following cladistic principles is evaluated: 1) only synapomorphies provide evidence for cladistic relationships, 2) *ad hoc* hypotheses are to be minimized in the generation of cladistic hypotheses, and 3) data used in the inference of cladistic relationships must be independent of each other. To be consistent with these principles, MRP analysis must 1) be based on source trees that were generated using cladistic analyses of character data, 2) weight the input data to account for the relative support for individual nodes on source trees and to eliminate inappropriate biases associated with variation in tree size, 3) be based on source trees with high consistency indices, and 4) be based on source trees that provide independent evidence for relationships. Achieving these criteria is extremely difficult, and all published MRP analyses fail to meet one or more of these conditions. Although MRP supertrees might be justified on pragmatic grounds, these trees should be considered a heuristic synthesis of hierarchical information, rather than a rigorous phylogenetic analysis of the included taxa.

**Keywords:** character data; cladistic analysis; cladistic principles; data independence; matrix representation with parsimony; parsimony; supertree construction

## 1. Introduction

Matrix representation with parsimony (MRP; Baum, 1992; Ragan, 1992) uses additive binary coding (Farris *et al.*, 1970) and parsimony analysis to construct one or more supertrees based on the hierarchical information in two or more overlapping cladograms, phylogenetic trees, or taxonomies (the source trees). In standard MRP, each node or component (*sensu* Wilkinson, 1994) on each source tree is represented by a binary “matrix element” (Baum and Ragan, 1993): terminal taxa in a particular component on a particular source tree are scored as 1 and all other taxa on that tree are scored as 0 for that component. Taxa in the matrix that are not present on that particular tree are scored as ?. Parsimony analysis of the matrix results in one or more shortest supertrees, which are rooted using an all-zero outgroup. Various modifications to the method proposed by Baum and Ragan have been proposed and MRP is but one of several matrix representation methods (for reviews, see Baum and Ragan, 2004; Wilkinson *et al.*, 2004).

MRP is a methodologically explicit means of making phylogenetic statements about sets of taxa that have not been included in a single character-based phylogenetic analysis (Sanderson *et al.*, 1998). Because of its ability to generate supertrees based on incongruent source trees, source trees with poor taxonomic overlap, and source trees that are based on different types of data analyzed using different methods, MRP has become a popular means for assembling large phylogenetic trees (e.g., Purvis, 1995; Bininda-Emonds *et al.*, 1999; Liu *et al.*, 2001; Jones *et al.*, 2002; Kennedy and Page, 2002; Pisani *et al.*, 2002; Salamin *et al.*, 2002).

MRP is similar methodologically to cladistic analysis of primary character data because of its use of a data matrix and parsimony analysis to generate supertrees, and the method has been compared favorably with analyses of primary evidence using total evidence methods (Bininda-Emonds *et al.*, 1999; Bininda-Emonds and Sanderson, 2001). However, MRP differs from standard cladistic analyses in being one step removed from the original character data. Tree topologies, rather than primary character data, are used to generate the element matrix that the supertree is based on. This loss of contact with the primary data and its consequences have resulted in strong criticism of the method (Rodrigo, 1993, 1996; Springer and de Jong, 2001; Wilkinson *et al.*, 2001; Gatesy *et al.*, 2002; Goloboff and Pol, 2002; Gatesy and Springer, 2004; Ross and Rodrigo, 2004). The suitability of MRP as a phylogenetic method has also been questioned because, although it is claimed to have character congruence-like properties (Bininda-Emonds *et al.*, 1999), the supertrees that result from MRP analyses of source trees based on partitioned data sets are not always the same as those generated using a total evidence analysis of the same

primary data (Pisani and Wilkinson, 2002). As a result, the latter authors considered MRP more similar to taxonomic congruence and concluded that it lacked a convincing theoretical justification.

In this contribution, I take a theoretical approach to the question of whether MRP is a valid phylogenetic method. In particular, I consider the more limited issue of the degree to which MRP is consistent with fundamental cladistic principles, and whether there are specific steps that a researcher can take in an MRP analysis to maximize the fit with those principles. MRP is amenable to such an analysis because of its methodological similarities to cladistic analysis of character data. For the purposes of this discussion, the following cladistic principles are assumed: 1) only synapomorphies provide evidence of cladistic relationship, 2) *ad hoc* hypotheses are to be minimized in the generation of cladistic hypotheses of relationship, and 3) data used to generate cladistic hypotheses must be independent of one other. In the sections that follow, I discuss the degree to which MRP analysis conforms to each of these principles. If MRP can be designed so that it is consistent with these principles, it would provide a theoretical justification for the method. This discussion pertains primarily to the original method proposed by Baum (1992) and Ragan (1992), but with some consideration of how proposed modifications might, or might not, make MRP more consistent with cladistic principles.

## 2. Evidential support for clades

The fundamental principle of phylogenetic systematics (cladistics) is that only synapomorphies (shared derived characters) provide evidence for phylogenetic relationships among taxa (Hennig, 1966); symplesiomorphies (shared primitive characters) or raw similarity do not provide valid evidence for such relationships. In cladistic analysis, taxa are scored for a set of cladistically informative characters, characters that potentially separate taxa into groups along internal branches of an unrooted tree. When the shortest unrooted trees found during parsimony analysis are rooted subsequently, the synapomorphic character states of those characters constitute the evidence for the clades that result.

Matrix elements in MRP fulfill the operational role of characters in standard parsimony analyses. A matrix element differs from a character in that, whereas characters are attributes of organisms, a matrix element refers to a component on a source tree and is therefore a topological unit (Baum and Ragan, 1993) or a membership criterion (Bininda-Emonds and Bryant, 1998). The fact that matrix elements do not represent synapomorphic evidence for relationships directly might suggest that MRP is inconsistent

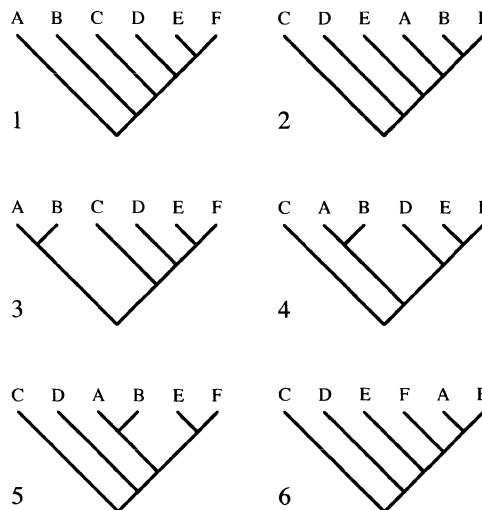
*Table 1.* Data matrix for the trees in Figure 1 (see text and legend for Figure 1).

| Taxon | Character |   |   |   |   |   |   |   |
|-------|-----------|---|---|---|---|---|---|---|
|       | 1         | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| A     | 0         | 0 | 0 | 0 | 1 | 1 | 1 | 0 |
| B     | 1         | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| C     | 1         | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| D     | 1         | 1 | 1 | 0 | 1 | 0 | 0 | 0 |
| E     | 1         | 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| F     | 1         | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

with this basic cladistic principle; however, matrix elements have the potential to act as proxies for synapomorphies if the source trees are based on primary character evidence that was analyzed cladistically. The components on those trees are based on this character evidence and therefore this evidence is represented indirectly in the matrix elements.

Consider the simple hypothetical example in Table 1 and Figure 1. Table 1 includes eight characters scored for six taxa; characters that support the monophyly of the entire group are not included for simplicity. The data are partitioned into two groups (characters 1–4 and characters 5–8). Trees 1 and 2 in Figure 1 are the most parsimonious trees based on characters 1–4 and characters 5–8, respectively. There is no homoplasy in either partition, and each internal node on each tree is supported by one synapomorphy; as a result, the consistency index (CI) of each character and of each tree is 1.0. When the partitions are combined into and analyzed as one data matrix, four equally most parsimonious trees result (trees 3–6). If, instead, trees 1 and 2 are used as source trees in MRP analysis, the element matrix is identical to the combined character matrix (Table 1), and the result, therefore, is four equally most parsimonious supertrees that are identical to the four trees generated using the primary character data (trees 3–6). In this simple example, the element matrix acts as an exact proxy for the original character data. There is an equivalence between the partitioned character data and the components on the two source trees (see also Williams 1994, 1996), and, thus, MRP is consistent in this instance with the principle that cladistic relationships are based on synapomorphic evidence. However, empirical data sets will deviate, often considerably, from this admittedly contrived example, thereby introducing biases in the representation of synapomorphic evidence.

MRP analyses of typical source trees result in differential weighting of the primary synapomorphic evidence. Each component or node on each source tree has the same weight in a standard MRP analysis. Invariably, however, the character support for the components on each source tree (as measured by bootstrap values, Bremer support, or the number of



**Figure 1.** Most parsimonious trees, source trees, and supertrees based on the data matrix in Table 1. Trees 1 and 2 are the most parsimonious topologies based on characters 1–4 and characters 5–8, respectively, in Table 1. Trees 3 through 6 are the four equally most parsimonious trees that result from a total evidence analysis of the eight characters in Table 1. Trees 3 through 6 are also the four equally most parsimonious supertrees that result from an MRP analysis of the element matrix based on the topologies of trees 1 and 2. All four of these supertrees contain the novel clade AB. Example modified from Goloboff and Pol (2002; figure 1).

synapomorphies) varies. As a result, individual synapomorphies of well-supported nodes will have less influence on the topology of the supertree than synapomorphies of less well-supported nodes.

Because matrix elements represent nodes on the source trees, source trees that include more taxa will usually have more nodes, and therefore be represented by more matrix elements in the MRP analysis; as a result, these trees will have a relatively larger effect on the topology of the MRP tree. All things being equal, this bias towards larger trees seems justified because larger trees contain more hierarchical information regarding phylogenetic relationships (see Bininda-Emonds and Bryant, 1998). However, trees with more nodes are not based necessarily on data matrices that include more informative character evidence. As a result, differences in the sizes of the source trees can also introduce differential weighting of the primary character evidence.

Differential weighting *per se* is not contrary to cladistic principles (Farris, 1983). In fact, weighting is unavoidable. The basic decision to include or exclude characters in a parsimony analysis is a weighting criterion. However, the weighting in MRP associated with its failure to

consider differential support for nodes and with any inappropriate influence of differences in the size of trees seems inconsistent with cladistic principles because it is not based on any empirical assessment of the evidence, but involves instead an inherent bias in the method. As a result, weighting that eliminates these biases in the basic methodology would make MRP more consistent with cladistic principles; however, preliminary analyses suggest that finding appropriate weighting schemes can be elusive (Bininda-Emonds and Bryant, 1998) or ineffective (example 3 in Wilkinson *et al.*, 2001).

Although MRP analyses of source trees that are based on cladistic analyses of character data can be consistent with cladistic principles, MRP analyses that include source trees that were not generated using such methods will not be completely consistent with those principles because they include matrix elements that need not be based on synapomorphic evidence. One might argue that, although these source trees are not necessarily based on synapomorphies, their matrix elements could still act as proxies for the phylogenetic evidence supporting the nodes on those trees. Unfortunately, the relationship between this phylogenetic evidence and actual synapomorphies would often be nebulous. Also, because these trees are not necessarily supported by synapomorphic evidence (e.g., trees based on some distance methods), these trees might not be up of clades. Source trees that are based on published taxonomies are especially problematic because they might not be based on empirical phylogenetic analyses of any kind. Thus, although the ability of MRP to combine hierarchical data based on a variety of methods and sources is one of its appeals (Bininda-Emonds and Sanderson, 2001), such analyses will almost certainly not be fully consistent with the fundamental principles of cladistic analysis.

### 3. Parsimony analysis in MRP

In standard cladistic analysis, tree choice is based on the principle of parsimony: one prefers the cladistic hypothesis or hypotheses that minimize the requirement for *ad hoc* hypotheses of homoplasy in the character evidence (Farris, 1983). Putative synapomorphies (“primary homologues”; *sensu* de Pinna, 1991) often provide contradictory evidence for cladistic relationships and application of the principle of parsimony results in the acceptance of the largest body of that evidence. The hypotheses that certain putative synapomorphies are homoplasies are erected solely to preserve the most parsimonious tree; the evidence these incongruent synapomorphies provide for alternative hypotheses of relationship remains. In combination with additional evidence for alternative hypotheses, they have the potential

to overturn the accepted hypothesis because it would no longer be the most parsimonious explanation for all available character evidence.

The method by which MRP resolves incompatibility among the topologies of the source trees is equivalent operationally to that used in character-based cladistic analysis. Parsimony is used to minimize the degree to which one is forced to reinterpret the hierarchical structure of the source trees represented by the matrix elements. But how should we reinterpret hierarchical information that is incongruent with the supertree? Rodrigo (1993) argued that the “homoplasy” associated with incompatible source tree topologies has no obvious “biological meaning” (see also Slowinski and Page, 1999; Burleigh *et al.*, 2004; Cotton and Page, 2004; Gatesy and Springer, 2004; Ross and Rodrigo, 2004). Wilkinson *et al.* (2001) saw no reason to prefer the topology of one source tree to that of any other, and concluded that conflict resolution in MRP was no more than a methodological artifact. From a cladistic perspective, it seems appropriate to reinterpret the hierarchical information in the source trees that is incongruent with that in the supertrees, rather than considering it simply a necessary by-product of conflict resolution with no biological meaning. Because the hierarchical information in source trees is ideally a reflection of congruence among cladistic character data, incongruent hierarchical information in the source trees should be explained in terms of the character data that the source trees are based on.

In character-based parsimony analysis, there are two ways of interpreting incongruence in the character data: convergence and reversal. In the first instance, transformations to the apomorphic character state are interpreted as occurring independently in subsets of the group of taxa with that state; in the second instance, there is a secondary reverse transformation to the plesiomorphic character state. The analogous situation in MRP to convergence involves the breaking apart of clades on source trees into two or more parts on the resulting supertrees; in the analogous situation to reversal, two or more taxa are clustered together on the supertree because they were not part of a particular clade on one or more of the source trees. The appropriateness of both of these results of MRP has been questioned. Goloboff and Pol (2002) argued that there was no justification for breaking apart clades on source trees into two or more parts, and various authors (Bininda-Emonds and Bryant, 1998; Goloboff and Pol, 2002) have questioned whether reversals should be allowed in MRP. Reversals are permitted in parsimony analysis of character data because the secondary loss of a synapomorphy within the clade it supports can act in turn as a synapomorphy at that less-inclusive level. By contrast, the 0s in the element matrix in MRP analysis represent lack of membership in a clade on a source tree (Bininda-Emonds and Bryant, 1998), a seemingly inappropriate basis for

clustering taxa on the supertree. Although prohibiting reversals had little effect on the results in the simulations of Bininda-Emonds and Sanderson (2001), irreversible matrix elements produced a supertree that resembled the total evidence tree more closely in an analysis of grasses (Salamin *et al.*, 2001).

The reinterpretation of hierarchical information in MRP analysis can be considered equivalent to convergence or reversal in the primary character evidence to the extent that the matrix elements act as proxies for putative synapomorphies. This interpretation is most reasonable when the matrix elements represent clades that are supported by unique synapomorphies (i.e., characters with a CI of 1.0 on source trees). In the example in Figure 1, where each node on each of the two source trees (trees 1 and 2) is supported by one unique synapomorphy (CI = 1.0), the matrix elements act as exact proxies for the primary character evidence, and the reinterpretation of the hierarchical information in the source trees is completely consistent with homoplasy in the character data in the combined character matrix. The breaking apart of clades occurring on the source trees in the resulting supertrees can be associated directly with the *ad hoc* reinterpretation of unique synapomorphies as instances of convergence. Also, the presence of clades on the supertrees that are supported by 0s in the element matrix can be associated directly with instances of reversal in the synapomorphies that support the nodes those matrix elements are based on. Thus, the arguments by Goloboff and Pol (2002), the source of the example in Figure 1, that the reinterpretation of hierarchical information by MRP is necessarily inappropriate are incorrect. However, in instances where the CI of source trees is less than 1.0, the hierarchical information in the matrix elements will no longer match that in the primary character data exactly. Almost invariably, source trees based on empirical data sets have CIs of less than 1.0, and often many clades on those source trees are supported only by characters with CIs below 1.0. In these instances, the breaking apart of clades on those source trees by MRP cannot be associated unambiguously with convergence and reversal in particular characters in the data matrix upon which those source trees are based.

If the CI of the source trees is less than 1.0, not all the character information used to generate the source trees is utilized in the generation of the supertree(s). Character evidence in the original data matrices that is incongruent (homoplastic) with the source trees will have no influence on the MRP analysis because that evidence is not represented by the nodes on those trees. The matrix elements represent the hierarchical information in the sources trees (primary signal; *sensu* Pisani and Wilkinson, 2002); character evidence in the original cladistic analyses that supports alternative hypotheses of relationship (subsignals; *sensu* Pisani and Wilkinson, 2002)

has no input into the generation of supertrees using MRP. Bininda-Emonds *et al.* (1999:146) argued that “MRP is essentially a parsimony analysis of the different phylogenetic signals within each data set stripped of any confounding noise (i.e. homoplasy).” However, this characteristic of MRP is problematic because, unlike total evidence analyses where character matrices are combined, the subsignals in those character matrices have no opportunity to interact and result in cladistic patterns in the supertree that were not present in the individual source trees. Proponents of MRP see this loss of character information as a necessary tradeoff in an analysis that is able to combine all possible sources of phylogenetic information (e.g., Bininda-Emonds *et al.*, 2002). Based on their simulations, Bininda-Emonds and Sanderson (2001) argued that this tradeoff is relatively minor. However, the presence of incongruent source trees implies that at least one of those trees is incorrect, and therefore that, contrary to the above quote, at least some characters that would be interpreted as homoplasies and not homologies in a total evidence analysis are represented in the MRP element matrix (see Pisani and Wilkinson, 2002). Thus, given that the CIs of source trees are almost always less than 1.0, MRP is expected to yield a more or less incomplete and biased representation of the phylogenetic information in the original character matrices.

The above discussion suggests that as the amount of homoplasy increases in the character matrices used to generate the source trees, the ability of MRP to represent that character evidence adequately will decrease. In other words, the relative amount of character evidence represented by subsignals that cannot contribute to the element matrix will increase. Pisani and Wilkinson (2002) provided an example in which MRP analysis of two source trees (each based on seven characters and containing the same four taxa) resulted in a supertree that was different from the tree that resulted from a total evidence analysis of the combined data matrix. In their example, the original partitioned data matrices each had a relatively large amount of homoplasy (three of seven characters were incongruent with each of the source trees). Subsignals in the partitioned data matrices provided the primary signal in the total evidence tree. This example suggests that the CIs of the source trees should provide at least a rough guide to the degree to which an element matrix in MRP represents the primary character evidence. This relationship should be explored in detail using simulation studies.

MRP can produce clades that do not occur on any of the source trees because 1) the source trees can have different taxa and the resulting supertree can include taxon combinations that are not present on any one source tree (Sanderson *et al.*, 1998), and 2) the parsimony analysis can resolve incongruence among source trees by forming novel clades (Bininda-Emonds and Bryant, 1998; Pisani and Wilkinson, 2002). Only the latter is

relevant to this discussion. Although Bininda-Emonds *et al.* (1999) viewed the generation of novel clades as a character congruence-like property of MRP, other authors (e.g., Gatesy and Springer, 2004) have viewed this behavior as a flaw in the method. Pisani and Wilkinson (2002) argued that, unlike total evidence analyses, the generation of novel clades in MRP does not involve character congruence because there is no interaction among subsignals in the initially separate data partitions. They concluded that novel clades generated in MRP analyses lack any known justification. Goloboff and Pol (2002) claimed that the novel clades produced by MRP are “spurious” because they necessarily contradict the topology of one or more of the source trees (i.e., the pattern of the supertrees deviates from the strict consensus solution).

The example in Figure 1 illustrates that the creation of novel clades in MRP can be justified from a cladistic perspective in some instances. Thus, if one accepts the criterion of maximum parsimony, the occurrence of the novel clade AB on each of the four supertrees (trees 3–6) is not problematic or “spurious.” In instances where the matrix elements act as exact proxies for the character evidence, the generation of novel clades can be associated directly with overall character congruence. Although no individual character in the matrix is congruent with clade AB, the occurrence of this clade is the most parsimonious interpretation of all the data. However, the association of novel clades with character congruence will pertain fully only when the CIs of the source trees are 1.0 and the character support for individual nodes is identical. If the CIs are less than 1.0, the matrix elements no longer act as exact proxies for the character evidence because subsignals that exist in the original character matrices are not represented. Also, as noted previously, differences in character support for nodes are not represented in an unweighted element matrix (although this second problem is potentially correctable through weighting; see Bininda-Emonds and Sanderson, 2001). Both of these situations reduce the character congruence-like properties of the MRP analysis. With typical source trees, much of the character congruence-like behavior of MRP might not pertain to the occurrence of novel clades, but might be restricted primarily to situations where the topologies of two or more source trees agree with one another, as in instances where those topologies are favored over incongruent topologies on other source trees included in the analysis.

The generation of novel clades in MRP might be an artifact of the method in instances where it cannot be associated with character congruence. This artifact might result from interaction among aspects of the topologies of incongruent source trees (e.g., shapes) that are not associated directly with the underlying character support for those source trees. In a series of examples designed to illustrate how MRP resolves conflicts

between two source trees, Wilkinson *et al.* (2001) suggested that the method seems to have a bias toward the crownward position of taxa and also seems to favor the topology of asymmetrical trees (for other properties / biases of MRP, see Wilkinson *et al.*, 2004). These biases seem related and associated with the fact that taxa in crownward positions on asymmetrical trees will tend to have more matrix elements that support that position than taxa in a more basal position on more symmetrical trees. Example 3 in Wilkinson *et al.* (2001) suggests that this shape bias cannot necessarily be overcome by weighting matrix elements based on the character evidence for those nodes (e.g., bootstrap values). The generation of novel clades based solely on the shapes of source trees seems inconsistent with the principles of cladistic analysis.

#### 4. Independence among matrix elements

For cladistic analysis to generate an unbiased analysis of the character evidence, the individual characters in the data matrix must represent independent sources of evidence of relationship. If two characters are in fact two separate representations of the same feature, inclusion of both characters in the analysis provides unjustified weight to that feature.

In MRP analysis, the source trees should be based on independent sources of character evidence. Source trees based on morphology and on molecular analyses of several different genes would be one example. However, when a series of source trees are culled from the literature for MRP analysis, there can be considerable overlap in the character evidence upon which those trees are based (see figure 3 in Gatesy *et al.*, 2002). For example, series of morphological cladistic analyses for a set of taxa often rely on some of the same character evidence. The problem will be exacerbated further if taxonomies that are based in whole or in part on those phylogenetic analyses are also used as the basis for source trees in the same analysis.

Bininda-Emonds *et al.* (2002) admitted that some lack of independence is probably inevitable in MRP analyses despite conscious attempts to minimize it (e.g., Purvis 1995; Bininda-Emonds *et al.*, 1999; Liu *et al.*, 2001; Jones *et al.*, 2002). For example, in the analysis in which Liu *et al.* (2001) generated a mammal supertree, there is redundancy in both the morphological characters and some of the molecular sequences upon which the source trees are based (Springer and de Jong, 2001; Gatesy *et al.*, 2002). Bininda-Emonds *et al.* (2002, 2004) acknowledged that lack of independence is a potential problem for MRP, but suggested that interaction between repeated data and other data in the source analyses, together with the fact that

employing different assumptions and methods with the same character evidence can produce different phylogenetic hypotheses, might alleviate those concerns. From a theoretical point of view, however, lack of independence among the evidence for individual source trees is a serious problem that violates a fundamental principle of cladistic analysis.

Sets of matrix elements lack independence in a second sense; all matrix elements from a single source tree are necessarily congruent (Bininda-Emonds *et al.*, 1999). In effect, a set of matrix elements from one source tree is equivalent methodologically to an additive binary coded character-state tree in a character-based cladistic analysis. This clique of matrix elements can be considered consistent with cladistic principles to the degree that it represents the hierarchical information in the primary character data. The character evidence upon which the source tree was based provides the justification for the “transformation series” represented by the clique of matrix elements.

## 5. Conclusion

This discussion suggests that individual MRP analyses must possess the following properties to be consistent with cladistic principles: 1) they must be based on source trees that were generated using well-designed cladistic analyses, 2) matrix elements or sets of matrix elements should be weighted based on the relative character support for individual nodes on the source trees and to alleviate inappropriate biases associated with tree size, 3) the source trees should have high CIs (i.e., high congruence in the original character matrices), and 4) the source trees must be based on different sets of character evidence (e.g., morphological versus molecular data sets, or trees based on different genes) to guarantee independence among the matrix elements. Criteria one, three and four are the easiest for the investigator to control. Trees based on non-cladistic methods or taxonomies should be avoided (e.g., Pisani *et al.* 2002) and cladistically generated source trees should be chosen with care based on the above criteria. Regarding criterion two, finding weighting methods that alleviate all the biases in MRP has been elusive to date (Bininda-Emonds and Bryant, 1998). However, simulations that suggest that weighted MRP analyses can produce supertrees that are very similar to model phylogenies and trees produced using total evidence methods (Bininda-Emonds and Sanderson, 2001) are encouraging. Clearly, more analysis is needed. The importance of criterion three cannot be overemphasized. One of the most serious inherent drawbacks of MRP analysis from a cladistic perspective is the loss of character information in the generation of the matrix elements based on most source trees; only the

primary signal is represented. Subsignals that have the potential to result in new or different cladistic relationships in total evidence studies are lost in MRP analysis. Given that, in some instances, the appearance of novel clades on MRP supertrees might be an artifact of the method that has nothing to do with the primary character evidence, Pisani and Wilkinson's (2002) suggestion that they be flagged on supertrees seems appropriate.

The analysis in this manuscript cannot be extended easily to the evaluation of other methods of supertree construction. The mechanical similarities between MRP and the cladistic analysis of character data make MRP especially amenable to analysis from a cladistic perspective. Because other supertree methods lack this methodological similarity, analysis from a cladistic perspective will be difficult and might not even be appropriate.

One might argue that the approach taken in this manuscript is too restrictive in that, by evaluating MRP within a strictly cladistic context, it ignores arguments for the use of alternative phylogenetic methods and the advantages of the ability of MRP to produce supertrees using source trees generated by different phylogenetic methods, often using data that cannot be combined (e.g., DNA-DNA hybridization data analyzed using distance methods and morphological characters evaluated using parsimony-based cladistic analysis). Thus, to realize the full potential of MRP, one might want to take a more permissive approach whereby trees obtained using only well-designed phylogenetic methods (cladistic and otherwise) are used in MRP analyses. After all, by limiting the trees to only those based on cladistic methodology, there is a strong argument for avoiding MRP methods all together in favour of total evidence approaches (*sensu* Kluge, 1989) using the original character data. Consequences of a more permissive approach include the absence of a cladistic justification for the method, and, given the differing theoretical bases for various phylogenetic methods, the likelihood that finding another theoretical justification for MRP will be extremely difficult.

This discussion suggests that, to date, justification for MRP analyses of typical source trees must be based more on pragmatic concerns than on fundamental phylogenetic principles. Proponents of MRP have argued that this method provides a means of achieving greater taxonomic and evidential coverage of phylogenetic information than is possible using traditional phylogenetic methods (Sanderson *et al.*, 1998; Bininda-Emonds *et al.*, 2002). Incompatible types of evidential data can be accommodated and the problems of combining data matrices that have poor taxonomic overlap can be alleviated, at least to some degree. Large comprehensive phylogenetic trees provide a basis for broad ecological and evolutionary analyses that are not possible using phylogenies with only limited taxonomic coverage (Bininda-Emonds *et al.*, 2002; Gittleman *et al.*, 2004; Moore *et al.*, 2004).

Thus, trees that include all extant species of Primates (Purvis, 1995), Carnivora (Bininda-Emonds *et al.*, 1999), and Chiroptera (Jones *et al.*, 2002) are extremely desirable. However, it is important to realize not only that these MRP analyses violate one or more of the cladistic principles discussed in this chapter, but that currently they also lack any other sound justification in phylogenetic theory. As a result, these MRP supertrees should be considered heuristic syntheses of available hierarchical information, rather than the products of rigorous phylogenetic analysis. Unfortunately, because of their comprehensive taxonomic coverage, there will be a tendency to use those supertrees as the phylogenetic basis for evolutionary and ecological inferences that might be spurious because of biased representation of the underlying character evidence by MRP, or the inclusion of source trees that are not based on rigorous cladistic or other phylogenetic methods.

It might be best to consider MRP supertrees as interim pragmatic solutions to the problem of generating comprehensive phylogenies until computer technology, phylogenetic methodology, and the collection of primary character data advances to the point where more theoretically justifiable analyses of data sets with a broad taxonomic coverage are possible. Nonetheless, the inconsistent and often poor taxonomic sampling of primary character data in many taxonomic groups and the methodological incompatibility of some types of character data suggest that supertree analysis might have an important role in the construction of taxonomically comprehensive phylogenies for the foreseeable future (see Bininda-Emonds *et al.*, 2002).

## Acknowledgements

I would like to thank Olaf Bininda-Emonds for asking me to contribute both to this book and the symposium on supertree construction held at “Evolution 2003” at California State University at Chico in June 2003. I also thank Davide Pisani, David Williams and Olaf Bininda-Emonds for their very helpful comments on the submitted version of this chapter.

## References

- BAUM, B. R. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon* 41:3–10.
- BAUM, B. R. AND RAGAN, M. A. 1993. Reply to A.G. Rodrigo’s “A comment on Baum’s method for combining phylogenetic trees”. *Taxon* 42:637–640.
- BAUM, B. R. AND RAGAN, M. A. 2004. The MRP method. In O. R. P. Bininda-Emonds (ed). *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 17–34. Kluwer Academic, Dordrecht, the Netherlands.

- BININDA-EMONDS, O. R. P. AND BRYANT, H. N. 1998. Properties of matrix representation with parsimony analyses. *Systematic Biology* 47:497–508.
- BININDA-EMONDS, O. R. P., GITTLEMAN, J. L., AND PURVIS, A. 1999. Building large trees by combining phylogenetic information: a complete phylogeny of the extant Carnivora (Mammalia). *Biological Reviews* 74:143–175.
- BININDA-EMONDS, O. R. P., GITTLEMAN, J. L., AND STEEL, M. A. 2002. The (super)tree of life: procedures, problems, and prospects. *Annual Review of Ecology and Systematics* 33:265–289.
- BININDA-EMONDS, O. R. P., JONES, K. E., PRICE, S. A., CARDILLO, M., GRENYER, R., AND PURVIS, A. 2004. Garbage in, garbage out: data issues in supertree construction. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 267–280. Kluwer Academic, Dordrecht, the Netherlands.
- BININDA-EMONDS, O. R. P., AND SANDERSON, M. J. 2001. Assessment of the accuracy of matrix representation with parsimony supertree construction. *Systematic Biology* 50:565–579.
- BURLEIGH, J. G., EULENSTEIN, O., FERNÁNDEZ-BACA, D., AND SANDERSON, M. J. 2004. MRF supertrees. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 65–85. Kluwer Academic, Dordrecht, the Netherlands.
- COTTON, J. A. AND PAGE, R. D. M. 2004. Tangled trees from molecular markers: reconciling conflict between phylogenies to build molecular supertrees. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 107–125. Kluwer Academic, Dordrecht, the Netherlands.
- DE PINNA, M. C. C. 1991. Concepts and tests of homology in the cladistic paradigm. *Cladistics* 7:367–394.
- FARRIS, J. S. 1983. The logical basis of phylogenetic analysis. In N. I. Platnick and V. A. Funk (eds), *Advances in Cladistics*, volume 2, pp. 7–36, Columbia University Press, New York.
- FARRIS, J. S., KLUGE, A. G., AND ECKHARDT, M. J. 1970. A numerical approach to phylogenetic systematics. *Systematic Zoology* 19:172–191.
- GATESY, J., MATTHEE, C., DESALLE, R., AND HAYASHI, C. 2002. Resolution of a supertree / supermatrix paradox. *Systematic Biology* 51:652–664.
- GATESY, J. AND SPRINGER, M. S. 2004. A critique of matrix representation with parsimony supertrees. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 369–388. Kluwer Academic, Dordrecht, the Netherlands.
- GITTLEMAN, J. L., JONES, K. E., AND PRICE, S. A. 2004. Supertrees: using complete phylogenies in comparative biology. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 439–460. Kluwer Academic, Dordrecht, the Netherlands.
- GOLOBOFF, P. A. AND POL, D. 2002. Semi-strict supertrees. *Cladistics* 18:514–525.
- HENNIG, W. 1966. *Phylogenetic Systematics*. University of Illinois Press, Urbana.
- JONES, K. E., PURVIS, A., MACLARNON, A., BININDA-EMONDS, O. R. P., AND SIMMONS, N. B. 2002. A phylogenetic supertree of the bats (Mammalia: Chiroptera). *Biological Reviews* 77:223–259.
- KENNEDY, M. AND PAGE, R. D. M. 2002. Seabird supertrees: combining partial estimates of procellariiform phylogeny. *The Auk* 119:88–108.
- KLUGE, A. J. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). *Systematic Zoology* 38:7–25.

- LIU, F.-G. R., MIYAMOTO, M. M., FREIRE, N. P., ONG, P. Q., TENNANT, M. R., YOUNG, T. S., AND GUGEL, K. F. 2001. Molecular and morphological supertrees for eutherian (placental) mammals. *Science* 291:1786–1789.
- MOORE, B. R., CHAN, K. M. A., AND DONOGHUE, M. J. 2004. Detecting diversification rate variation in supertrees. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 487–533. Kluwer Academic, Dordrecht, the Netherlands.
- PISANI, D. AND WILKINSON, M. 2002. Matrix representation with parsimony, taxonomic congruence, and total evidence. *Systematic Biology* 51:151–155.
- PISANI, D., YATES, A. M., LANGER, M. C., AND BENTON, M. J. 2002. A genus-level supertree of the Dinosauria. *Proceedings of the Royal Society of London B* 269:915–921.
- PURVIS, A. 1995. A composite estimate of primate phylogeny. *Philosophical Transactions of the Royal Society of London B* 348:405–421.
- RAGAN, M. A. 1992. Phylogenetic inference based on matrix representation of trees. *Molecular Phylogenetics and Evolution* 1:53–58.
- RODRIGO, A. G. 1993. A comment on Baum's method for combining phylogenetic trees. *Taxon* 42:631–636.
- RODRIGO, A. G. 1996. On combining cladograms. *Taxon* 45:267–274.
- ROSS, H. A. AND RODRIGO, A. G. 2004. An assessment of matrix representation with compatibility in supertree construction. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 35–63. Kluwer Academic, Dordrecht, the Netherlands.
- SALAMIN, N., HODKINSON, T. R., AND SAVOLAINEN, V. 2002. Building supertrees: an empirical assessment using the grass family (Poaceae). *Systematic Biology* 51:134–150.
- SANDERSON, M. J., PURVIS, A., AND HENZE, C. 1998. Phylogenetic supertrees: assembling the trees of life. *Trends in Ecology and Evolution* 13:105–109.
- SLOWINSKI, J. B. AND PAGE, R. D. M. 1999. How should species phylogenies be inferred from sequence data? *Systematic Biology* 48:814–825.
- SPRINGER, M. S. AND DE JONG, W. W. 2001. Which mammalian supertree to bark up? *Science* 291:1709–1711.
- WILKINSON, M. 1994. Common cladistic information and its consensus representation; reduced Adams and reduced cladistic consensus trees and profiles. *Systematic Biology* 43:343–368.
- WILKINSON, M., THORLEY, J. L., LITTLEWOOD, D. T. J., AND BRAY, R. A. 2001. Towards a phylogenetic supertree of Platyhelminthes? In D. T. J. Littlewood and R. A. Bray (eds), *Interrelationships of the Platyhelminthes*, pp. 292–301. Chapman-Hall, London.
- WILKINSON, M., THORLEY, J. L., PISANI, D., LAPOINTE, F.J., AND MCINERNEY, J. O. 2004. Some desiderata for liberal supertrees. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 227–246. Kluwer Academic, Dordrecht, the Netherlands.
- WILLIAMS, D. M. 1994. Combining trees and combining data. *Taxon* 43:449–453.
- WILLIAMS, D. M. 1996. Characters and cladograms. *Taxon* 45:275–283.

## Chapter 17

# A CRITIQUE OF MATRIX REPRESENTATION WITH PARSIMONY SUPERTREES

John Gatesy and Mark S. Springer

**Abstract:** Strict and semi-strict supertree construction methods can be used to summarize groups that are consistent with all source phylogenies. Other procedures, such as Matrix Representation with Parsimony (MRP), arbitrate conflicts among incompatible source trees, and can provide more topological resolution than strict and semi-strict methods. MRP has been used to construct most of the large supertrees that have been published to date. We review some of the inherent problems with MRP and other supertree methods, point out specific difficulties in previously published MRP-supertree analyses, question some of the possible advantages of supertrees, and suggest that supermatrix analyses of character data should provide the primary framework for comparative biology in the 21st century.

**Keywords:** character; computational efficiency; hidden support; non-independence; supermatrix

### 1. Introduction

Multiple data sets, often including characters from both morphological and molecular studies, are now available for examining evolutionary relationships for a variety of taxonomic groups. A major challenge in systematics is integrating these data into an informed, well-supported hypothesis of phylogeny. Against the backdrop of this challenge, supertree and supermatrix methods have emerged as competing strategies for combining different types of data (Sanderson *et al.*, 1998). Whereas supermatrix methods entail pooling character data before phylogenetic analysis (Kluge, 1989), supertree methods combine trees derived from

*Bininda-Emonds, O. R. P. (ed.) Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life, pp. 369–388. Computational Biology, volume 3 (Dress, A., series ed.). © 2004 Kluwer Academic Publishers.*

different data sets (Sanderson *et al.*, 1998). Strict and semi-strict supertree methods summarize groups that are not contradicted by individual source trees (e.g., Steel, 1992; Goloboff and Pol, 2002). Other procedures, such as Matrix Representation with Parsimony (MRP; Baum, 1992; Ragan, 1992), can arbitrate topological conflicts among source trees and, in some cases, resolve groups that are incompatible with all source trees (Bryant, 2004; and references therein).

MRP is a popular method that has been used to construct most of the large supertrees that have been published to date, including families / orders of mammals (Liu *et al.*, 2001), bats (Jones *et al.*, 2002), carnivores (Bininda-Emonds *et al.*, 1999), primates (Purvis, 1995a), seabirds (Kennedy and Page, 2002), grasses (Salamin *et al.*, 2002), and bacteria (Daubin *et al.*, 2001). Here, we review inherent problems in the MRP method as well as specific deficiencies in published MRP analyses, and question possible advantages of supertree analysis. Because MRP has been the method of choice for most large supertree analyses, we focus our critique on this procedure. However, some difficulties in MRP-supertree analysis apply to all supertree methods. Throughout this chapter, “MRP supertree” and the more general term “supertree” are used to distinguish between MRP-supertree analysis and supertree analysis in general.

## 2. Problems and limitations of MRP supertrees

In this section, we survey characteristics of published MRP-supertree analyses that are, in our view, problematic. We suggest solutions for some of these difficulties, but do not argue that these stopgap fixes necessarily justify the MRP-supertree approach. Note also that many of the arguments below apply to all supertree analyses and not just MRP-supertree analysis.

### 2.1 Non-independence

In principle, supertrees should be constructed from source phylogenies that are based on independent, non-overlapping data. Otherwise, some data inadvertently will be weighted more heavily than other data. This prerequisite for supertree construction requires careful inspection of the characters that are the basis for source phylogenies. In practice, and especially for cases involving complex and diverse data sets, this problem is not trivial. For example, in an MRP-supertree analysis of relationships among the families / orders of placental mammals, Liu *et al.* (2001:1789) recognized the importance of this problem and used “source phylogenies that were based on different character sets and/or were unique by >30% of their

OTUs." This screening procedure for identifying overlapping information was not effective in the sense that some data were still represented disproportionately. The 430 source phylogenies analyzed by Liu *et al.* (2001) included several with high degrees of overlap. As noted by Springer and de Jong (2001), a single transferrin immunology data set for bats was incorporated into five different source trees. Gatesy *et al.* (2002) documented even more pervasive problems for the cetartiodactyl data included in Liu *et al.* (2001).

Others have employed more stringent rules to ensure that source phylogenies are based on independent data (see Bininda-Emonds *et al.*, 2004). For example, Jones *et al.* (2002) used several criteria to minimize data duplication in constructing their MRP supertree for bats. These included using the most recent or most complete analysis of a data set and excluding previous supertrees that were themselves secondary rather than primary analyses. Jones *et al.* (2002) avoided much of the data duplication that plagued Liu *et al.*'s (2001) study, but some redundancy remained. Examples included overlapping mitochondrial rRNA sequences (from Hollar and Springer, 1997; Hoofer and Van Den Bussche, 2001; Murphy *et al.*, 2001; Van Den Bussche and Hoofer, 2001) and redundant morphological characters (from Novacek, 1980; Simmons and Geisler, 1998). In our view, all redundant source data are unacceptable and must be eliminated to avoid the non-independence problem.

Overlapping data can be avoided in supertree studies by devoting more attention to the compilation and analysis of primary data rather than relying exclusively on previously published source phylogenies. For example, mitochondrial 12S rRNA genes are among the most commonly sequenced genes for many mammalian groups. However, these sequences will seldom be found in a single data set that is optimized for the taxonomic problem of interest. Given that relevant data might be reported in different studies, supertree builders should compile the most comprehensive 12S rRNA data set possible, subject that matrix to primary character analysis, and then use the resultant tree as the single source phylogeny for 12S rRNA. For bats, this would require merging 12S rRNA data from several publications (e.g., Teeling *et al.*, 2000; Hoofer and Van Den Bussche, 2001; Murphy *et al.*, 2001; Van Den Bussche and Hoofer, 2001). Construction of this revised rRNA data set also would demand, minimally, reanalyzing the Murphy *et al.* (2001) data set, a concatenation of 18 gene fragments, following exclusion of the mitochondrial 12S rRNA gene from the concatenation. If such additional analyses of the primary data were executed, duplications of evidence in supertree analysis are unnecessary.

With this extra work, however, one of the primary bases for the utility of supertree analysis, namely the limited research effort and time required to

build a synthetic tree (Bininda-Emonds *et al.*, 2002), might be lost. The construction and analysis of many additional data sets could make MRP-supertree studies less efficient, and more time consuming, relative to a single simultaneous supermatrix analysis of all relevant data (see also Section 3.3). To date, most published MRP-supertree data sets have included duplications of character data, in contrast to published supermatrix data sets where reviewers generally have not tolerated such redundancies. Supertree analyses of wholly extinct clades, such as ornithischian dinosaurs (see Pisani *et al.*, 2002), are especially prone to duplications of character evidence. This is because fossils can be scored usually for only a limited set of gross anatomical traits that are recycled in many publications.

## 2.2 Quality control

Another issue facing supertree construction is quality control, both for primary data and for source phylogenies based on these data. Should informal cladograms that cannot be reconstructed from primary data matrices be incorporated into supertrees? Should taxonomies be included in supertrees? How about trees derived from review articles written by authorities? If the goal is to summarize previous opinions on the phylogeny of a group, these procedures might be defensible. If the goal is to build phylogenetic hypotheses that reflect primary data and analytically robust studies, hand-drawn cladograms, reviews, and taxonomies that lack support from primary data matrices should be eschewed (see also Section 3.1). We acknowledge that this quality control problem applies also to supermatrix studies. To date, however, many published MRP-supertree matrices have included source trees that are based on dubious analyses or data (e.g., Purvis, 1995a; Liu *et al.*, 2001; Jones *et al.*, 2002; Kennedy and Page, 2002), so we think that this is an important issue.

In view of the quality control problem, some workers have suggested that older topologies could be excluded from supertrees (e.g., Pisani *et al.*, 2002). For example, Jones *et al.* (2002) constructed an MRP supertree for bats using only source phylogenies for the years 1970–2000. The general approach of eliminating older topologies might be relatively effective for purging unwanted data, but it is not a substitute for detailed quality control. As recently as the late 1980s, for instance, the UPGMA method was used routinely to analyze DNA-DNA hybridization data (Sibley and Ahlquist, 1990, and references therein), but most modern workers have abandoned UPGMA, because the quirky assumptions of the method are rarely justified. Other DNA-DNA hybridization studies from the 1980s used more appropriate distance methods that did not assume rate uniformity among lineages. This type of specific information should be taken into account

when deciding which source phylogenies to incorporate into a supertree analysis.

Supertree researchers can be confronted also with multiple trees for the same data set. This can occur when different authors analyzed the same data set or when the same author presented trees based on different methods (e.g., likelihood, parsimony, neighbor joining). In these cases, how should researchers decide which source phylogenies to incorporate into supertree analysis? Jones *et al.* (2002) employed two strategies. When different authors analyzed the same data, the most recent reanalysis was taken as the source phylogeny. When the same authors presented more than one tree based on the same data or overlapping data, these trees were combined into a single source phylogeny using MRP (see also Bininda-Emonds *et al.*, 2004). In our view, this approach sacrifices prudence for convenience. For example, in their analysis of higher-level relationships among the orders of placental mammals, Madsen *et al.* (2001) presented both maximum likelihood (ML) and maximum parsimony (MP) trees for a segment of the *BRCA1* gene for 51 placentals and a marsupial outgroup. ML and MP identified the same four groups of placental orders in unrooted analyses, but results of rooted analyses were considerably different. ML rooted between Afrotheria and other placentals, whereas the MP root rendered Rodentia paraphyletic at the base of the placental tree. Subsequent analyses suggested long-branch attraction in the MP analysis. Given these results, combining the ML and MP trees into a single source phylogeny might be unwarranted. As for problems with overlapping data, it is incumbent on supertree researchers to evaluate source phylogenies judiciously before including them in a supertree analysis, even if this requires that additional analyses be performed with the original data.

### 2.3 Weighting nodes in source phylogenies

Weighting issues apply to both supermatrix and supertree methods. For example, one contentious issue in the case of supermatrices is whether or not to apply equal weights to molecular and morphological characters. A fundamental issue in supertree construction is the weighting of weakly versus strongly supported nodes in individual source phylogenies (Purvis, 1995b; Ronquist, 1996; Bininda-Emonds and Bryant, 1998; Sanderson *et al.*, 1998; Bininda-Emonds and Sanderson, 2001). Should we apply equal weights to two nodes in the same source phylogeny that have completely different levels of support? In our view, failure to recognize this problem betrays the information content of the original data (see Section 2.4), and several authors have suggested that differential weighting of nodes in source trees is beneficial for MRP-supertree analysis (Ronquist, 1996; Bininda-

Emonds and Sanderson, 2001). However, most of the weighting functions that have been offered are not applicable to supertree matrices that include trees based on discretely coded characters and trees based on distance data (e.g., the MRP-supertree data set of Kennedy and Page (2002) included topologies based on DNA sequences and a topology based on DNA-DNA hybridization distances).

Ronquist (1996) suggested that decay indices (Bremer, 1994) or bootstrap percentages (Felsenstein, 1985) might be used to differentially weight matrix elements (nodal characters) that compose an MRP data set (see also Sanderson *et al.*, 1998; Bininda-Emonds and Sanderson, 2001). The decay index, the difference in steps between the shortest trees without and with a particular group, requires exact counts of character state changes, but this is not an option for trees based on distance or ML analyses (Ronquist, 1996; Bininda-Emonds and Bryant, 1998). Another possibility is to weight nodes as a function of bootstrap support. One advantage of this method is that bootstrap values can be calculated for MP, ML, and some distance analyses, but the bootstrap approach also has shortcomings. For example, two clades might receive equal bootstrap scores although other measures indicate wide differences in character support (e.g., bootstrap scores of 100% for each clade, but decay indices of 4 versus 200 extra steps). This is because the upper limit for the bootstrap is capped at 100%, whereas differences in parsimony or likelihood scores between competing topologies have no upper limit. As more and more character evidence for a particular hypothesis accrues, support can go higher and higher (as it should) according to the decay index or likelihood-difference measures, but not according to the bootstrap (for examples, see Gatesy *et al.*, 1999; Lee and Hugall, 2003). Furthermore, bootstrap scores are not applicable to all types of systematic data. Bootstrap analyses entail resampling of characters with replacement and are undefined for data sets that are not composed of discretely coded characters (e.g., DNA-DNA hybridization or immunological distance matrices).

The use of branch lengths as indicators of support has been critiqued sharply (Farris *et al.*, 2001; for an alternative opinion, see Wilkinson *et al.*, 2003), and other possible weighting methods, such as T-PTP probabilities or the number of unique synapomorphies at a node, are strictly character-based (Bininda-Emonds and Bryant, 1998). Thus, if unequal weighting of matrix elements in MRP-supertree analysis is considered beneficial (Bininda-Emonds and Sanderson, 2001), this differential weighting necessarily will require exclusion of certain incompatible data sets from the supertree analysis. Alternatively, all data sets could be included, but differential weighting of matrix elements would not be applicable generally. Inclusiveness (see Section 3.1) and differential weighting do not seem to be

compatible concepts in the supertree paradigm, at least until novel supertree methods are developed.

Again, it is worth noting that one of the primary bases for the utility of supertree analysis is the limited research effort and time required to build a synthetic tree (Purvis, 1995a; Liu *et al.*, 2001; Sanderson *et al.*, 1998; Bininda-Emonds *et al.*, 2002). However, trees published by different researchers do not always utilize the same methods of analysis and support measures. Thus, if supertree researchers favor a common measure of support to differentially weight source trees, such as the bootstrap, a certain amount of primary data reanalysis would be beneficial. In other words, all source trees that do not have published bootstrap scores in the literature would require reanalysis. We suggest that this extra analysis and computation, especially time-consuming resampling routines such as ML bootstraps for small indecisive data sets, could make MRP-supertree analysis less efficient and more time consuming relative to a single simultaneous supermatrix analysis of all relevant data (see also Section 3.3).

## 2.4 Black-box problems

For trees that are constructed from primary character data using parsimony and likelihood methods, the mapping and contribution of each character to the resulting tree can be calculated, and there is a logical connection between implied evolutionary character transformations and the choice among competing phylogenetic hypotheses (Farris, 1983; de Queiroz and Poe, 2001). MRP supertrees lack these important properties (Baum, 1992; Ragan, 1992; Rodrigo, 1993, 1996; Slowinski and Page, 1999; Wilkinson *et al.*, 2001), and other methods for constructing supertrees from incompatible source trees, such as semi-strict (Goloboff and Pol, 2002), average consensus (Lapointe and Levesque, 2004), and MINCUTSUPERTREE (Semple and Steel 2000) also entail a loss of contact with the primary character data.

All published supertree methods misinterpret hidden character support and conflicts within different data sets that emerge in supermatrix analysis (Barrett *et al.*, 1991). This is because heterogeneities in phylogenetic signal within different data sets generally are screened out in supertree techniques that combine trees instead of characters. Perhaps most alarming is the possibility that novel systematic hypotheses can emerge through MRP analysis, although these hypotheses are not supported by combined analysis of the primary character data and are contradicted by all source trees that compose the supertree data set (Bininda Emonds and Bryant, 1998; Goloboff and Pol, 2002; Bryant, 2004). This effect contrasts with supermatrix analysis, where novel hypotheses are an expected outcome of summing individual data partitions, each of which contains heterogeneous

phylogenetic signals. Emergent clades in published MRP supertrees have been rare (Bininda-Emonds, 2003), but, even if no novel clades emerge in MRP-supertree analysis, resolution of conflicts among different source trees is difficult to interpret in this framework (Goloboff and Pol, 2002).

Bininda-Emonds and Sanderson (2001) noted that MRP is well-grounded in basic graph and network theory, but this connection holds only for single source trees. More importantly perhaps, MRP is not well-grounded in phylogenetic theory, in contrast to certain supertree methods that interpret source-tree incompatibility in terms of evolutionary events such as gene duplication, gene loss, and horizontal transfer (Maddison, 1997; Slowinski and Page, 1999; Cotton and Page, 2004). Slowinski and Page (1999:818) argued that MRP "is flawed because homoplasy in this context has no obvious biological meaning." At some level, MRP-supertree analysis is a systematic "black box" (Wilkinson *et al.*, 2001). Concepts such as convergence and reversal of MRP matrix elements are difficult to interpret (Rodrigo, 1993, 1996), and unique clades that emerge in supertree analysis are symptomatic of the MRP black box effect (Pisani and Wilkinson, 2002). Thus, even if there is no duplication of character data in an MRP-supertree study, as in an analysis of several independent gene trees (Baum, 1992), the resolution of incongruence among different source trees has not been justified adequately in the MRP framework (see Rodrigo, 1993, 1996; Slowinski and Page, 1999; Wilkinson *et al.*, 2001; Goloboff and Pol, 2002; Bryant, 2004).

Several authors have suggested that supertrees might be conservative phylogenetic hypotheses (Jones *et al.*, 2002; Kennedy and Page, 2002) in that they are unlikely to resolve phylogenetic groups that are incorrect and will simply show lack of resolution instead. Jones *et al.* (2002:247) pointed out that, "Because the supertree method depends on congruence of source trees to support clades, lack of information and significant disagreements among studies tend to be reflected as loss of resolution in the consensus supertree." However, as mentioned above, MRP-supertree data sets can support novel groups that conflict with all source trees and clades that are not implied by any combination of source trees (see Goloboff and Pol, 2002). Considering that these emergent groups are contradicted by all the separate analyses and/or implied by no combination of source trees, it would be difficult to argue that such clades are conservative phylogenetic hypotheses from a topological-congruence perspective. Pisani *et al.* (2002) argued that novel groups that conflict with all source trees in an MRP analysis have no empirical or logical basis, and should be flagged as being dubious. Because such groups are thought to be unjustified aberrations of the MRP method (see discussion in Bryant, 2004), we suggest that MRP supertrees should not necessarily be considered to be conservative

phylogenetic hypotheses. Truly conservative supertree methods do exist (e.g., strict and semi-strict supertree methods), but these techniques also ignore hidden character support in different data sets.

Unlike traditional reviews, MRP uses an explicit optimality criterion to summarize common phylogenetic patterns among published data sets, and this has been cited as a positive quality of this procedure (e.g., Sanderson *et al.*, 1998; Kennedy and Page, 2002; Pisani *et al.*, 2002). Unfortunately, if a particular optimality criterion has no logical basis, the explicitness of the optimality criterion is irrelevant. Wilkinson *et al.* (2001:300) went so far as to state that the black-box methodologies of MRP "... resolve conflicts without promoting any understanding . . . . In the light of the known behavior of MRP, it remains to be demonstrated that there is any convincing justification for this approach to supertree construction." Perhaps because there is no coherent basis for MRP that links primary data and hypothesized evolutionary events to the resulting supertree, proponents of this procedure have turned instead to inductive extrapolations from computer simulations for justification (see Section 3.5).

### **3. Previous arguments for the utility of MRP supertrees**

In this section, we survey proposed advantages of MRP-supertree analysis, and question the logic and content of these arguments. For each case, we present a counterargument from the supermatrix perspective, and suggest that the perceived benefits of MRP-supertree analysis are minimal. Again, many of the arguments below apply to all supertree analyses generally and not just MRP-supertree analysis.

#### **3.1 Utilization of more systematic data**

Numerous authors have argued that an advantage of supertree analysis is that systematic data that cannot be combined into a supermatrix can be analyzed in supertree analysis (Purvis, 1995a; Sanderson *et al.*, 1998; Bininda-Emonds *et al.*, 1999, 2002; Kennedy and Page, 2002). Supermatrices generally require discrete-character data, and specific hypotheses of character homology are crucial in this approach (see Patterson, 1982). Trees produced from distance data such as DNA-DNA hybridization comparisons, immunological reactions, and morphometric shape measures can be encoded directly into a supertree data set, but apparently cannot be utilized in a supermatrix study (Sanderson *et al.*, 1998). Likewise, the hierarchical information in classifications and informal morphological studies has been used to build MRP supertrees (e.g., Liu *et al.*, 2001; Jones *et al.*, 2002;

Kennedy and Page, 2002), but is uninformative in a supermatrix context. Thus, proponents of supertrees have argued repeatedly that their approach can employ even more systematic information than “total-evidence” supermatrix analyses (e.g., Bininda-Emonds *et al.*, 2002). But, what is the nature of the data sets that cannot be incorporated into a supermatrix?

New DNA-DNA hybridization and immunological studies are rare. Perusal of the major systematics and evolutionary-biology journals from the past ten years suggests that these methods are nearly defunct. This is not by chance. Both methods are costly and time consuming, often do not include all pairwise comparisons, and have been replaced by methods that yield more detailed descriptions of genetic data in the form of explicit hypotheses of character state homology among individual nucleotides.

DNA-DNA hybridization distances summarize information from all “single-copy” genes in the genome by assuming a consistent relationship between heteroduplex melting curves and overall DNA sequence divergence. The ability of this method to sort out orthologous from paralogous relationships of different gene duplicates has been questioned (Marks *et al.*, 1988). Perhaps more importantly, when DNA-DNA hybridization distances are combined with data from individual genes in a supertree analysis, there is at least some redundancy of genetic information (see Section 2.1). In a large supertree data set that includes many gene trees, there can be a large overlap of data with a DNA-DNA hybridization tree. This duplication of evidence is unnecessary. Simply excluding DNA-DNA hybridization data, as in the supermatrix approach, is to us more justified.

Other distance data are characterized by analogous problems. Relative to DNA or amino-acid sequence data, immunological distances have no apparent advantage. Immunological comparisons attempt to summarize differences between orthologous protein sequences by assuming a consistent, reliable relationship between the strength of immunological reaction and the amount of amino-acid sequence divergence. Much more precise measurements of amino-acid sequence divergences are derived now from direct sequencing of genes or proteins, and have made vague immunological distances obsolete. The systematic utility of morphometric distances has been questioned also, even by proponents of this paradigm (e.g., Bookstein, 1994). But, in contrast to DNA-DNA hybridization scores and immunological distances, morphometric shape can be coded into discrete states using different procedures (Swiderski *et al.*, 1998). Therefore, these data can be included in a supermatrix analysis as independent systematic evidence, so the ability to incorporate morphometric data does not necessarily distinguish supertree analysis from supermatrix analysis.

Taxonomic classifications, which are not based on explicit character analysis, have been utilized in several MRP-supertree studies (e.g., Purvis,

1995a; Bininda-Emonds *et al.*, 1999; Jones *et al.*, 2002). These summaries often contain all, or most, species in a particular taxonomic group. Such data sets that link all taxa in a supertree analysis are seen as especially beneficial (Bininda -Emonds and Sanderson, 2001), but the empirical basis for a particular classification might not be stated clearly. Many classifications do not result from explicit analyses of documented data matrices, but are based simply on the cumulative knowledge, or whims, of particular authorities (see Section 2.2). Some supertree proponents have argued that any information on phylogeny is better than no information (Purvis, 1995a), but we suggest that if no explicit character information exists for a particular taxon, then new character data from that taxon need to be collected. Until this data collection is completed, it would be prudent to exclude taxa without characters from phylogenetic analysis (see Section 3.4).

Among the wide range of commonly used phylogenetic evidence available to modern systematists, only partially redundant DNA-DNA hybridization data that contain dubious or untestable hypotheses of orthology, immunological distances that are obsolete, and classifications that have no clear empirical basis cannot be incorporated into supermatrices. In our view, this is not a great loss of systematic information. If a researcher wished to compare DNA-DNA hybridization results, immunological trees, and classifications with supermatrix results, this could be done simply by documenting compatible and incompatible groups in the different topologies.

### 3.2 Cost effectiveness

Several authors have argued that supertrees are cost effective relative to supermatrices (Sanderson *et al.*, 1998; Bininda-Emonds *et al.*, 1999; Kennedy and Page, 2002). This statement would seem uncontroversial because no novel character data were collected in most previously published supertree studies. The contribution of no new empirical data is cost effective generally in terms of laboratory supplies and work hours, but it is not clear that constructing a supertree from the literature is significantly less time consuming than constructing a supermatrix from the literature and genetic databases. If time is equated with money, the time it takes to complete a study is a cost-effectiveness issue as well. For example, Liu *et al.* (2001) apparently took over two full years to construct their MRP supertree of mammals that included 91 taxa and 1965 informative matrix elements (see Hoover, 2001). By contrast, Gatesy *et al.* (2002) constructed a cetartiodactyl supermatrix of 75 taxa and 14 124 informative characters (~37 000 characters total) in less than two months. Because there have been few attempts at making comprehensive supermatrices, it is difficult to judge presently how time consuming this approach will be. However, from our

experience, different authors have been very generous in sharing published sequences, alignments, and character matrices. Because published data sets can be transported rapidly by email, the construction of very large supermatrices should not be viewed as necessarily more problematic, time consuming, or costly than the construction of large MRP-supertree data sets. Only non-redundant, discrete characters are utilized in supermatrix analyses. Therefore, the amount of data analyzed in a supermatrix analysis for a particular group is more restricted than many published supertree analyses, in which all sorts of data have been considered relevant, and redundant sampling of characters was deemed to be acceptable (e.g., see Purvis, 1995a; Bininda-Emonds *et al.*, 1999; Liu *et al.*, 2001; Jones *et al.*, 2002; Kennedy and Page, 2002). Because time is saved by avoiding dubious data and by not analyzing the same data multiple times (see Sections 2.1, 2.2, and 3.1), supermatrix construction could be less time consuming and more cost effective relative to supertree construction in some cases.

### 3.3 Computational efficiency

It has been suggested that MRP-supertree data sets tend to be internally consistent because homoplasy in source trees is filtered out before MRP analysis (Bininda-Emonds and Sanderson, 2001), and that this could make MRP searches tractable for large numbers of taxa. However, the same heuristic parsimony algorithms utilized in MRP-supertree analysis are used in supermatrix analysis, and these algorithms are slowed generally by increases in the number of taxa. Given current implementations, complete branch swapping for large trees is time consuming and tedious, and examination of the few completed large-scale MRP-supertree studies suggests that heuristic parsimony searches of large supertree data sets have not been simple. Many published analyses required truncation of searches before branch swapping was completed (e.g., Bininda-Emonds *et al.*, 1999; Wojciechowski *et al.*, 2000; Jones *et al.*, 2002; Kennedy and Page, 2002; Pisani *et al.*, 2002; Salamin *et al.*, 2002) or elimination of problematic “wildcard” taxa (Liu *et al.*, 2001). As a result of conflicts among source trees and missing data, there were many thousands of minimum-length trees for these supertree data sets, and it is not clear whether tree searches were effective because branch swapping was terminated prematurely.

The majority of published supertrees have, in part, evaded computational difficulties by assuming the monophyly of certain taxa, so that the supertree analysis is compartmentalized (e.g., Bininda-Emonds *et al.*, 1998; Jones *et al.*, 2002). In this way, several smaller MRP parsimony searches can be executed, with the results of these separate analyses pasted together to form a complete supertree for the group of interest. This procedure breaks a large

systematic problem with many taxa into several smaller problems, but defeats the basic rationale for doing a single comprehensive supertree analysis in the first place. By assuming the monophyly of certain subgroups, supermatrix analyses become tractable for speciose clades as well (Wilkinson *et al.*, 2001); in fact, the combination of compatible subgroups has been used traditionally to create large synthetic trees over the past century. Some supertree proponents have labeled these methods as *ad hoc* (Sanderson *et al.*, 1998), but most large MRP analyses also have required *ad hoc* compartmentalization of searches to obtain publishable results.

Given that few supertree studies have been completed, that published MRP-supertree searches have been rough going, and that supermatrix analyses have been rare, we contend that no generality about the rapidity of MRP-supertree searches relative to supermatrix searches has emerged from the literature. The cetartiodactyl data set compiled by Gatesy *et al.* (2002), one of the larger published supermatrices in terms of data sets included, supported a single most parsimonious tree, and heuristic-parsimony searches swapped to completion very rapidly (~2.5 hours to do 1000 random taxon addition replicates with TBR branch swapping in PAUP\* v4.0b8; Swofford, 1998). By contrast, the MRP-supertree data set of Liu *et al.* (2001) is much less decisive, has many equally most parsimonious solutions, and requires much more time to do an analogous tree search on PAUP\* (~12 hours). We do not argue here that supermatrix searches always will be rapid relative to MRP-supertree searches, but the example above does show that some supermatrix searches can be executed rapidly.

Other supertree methods such as the strict and semi-strict approaches (Steel, 1992; Goloboff and Pol, 2002) are more promising in terms of analysis time. Given even moderate conflict among source trees, however, the semi-strict procedure will produce poorly resolved topologies, and the strict supertree method can be applied only to compatible source trees. Thus far, such procedures have not been used to build large synthetic supertrees from many source phylogenies. We contend that if the well-justified semi-strict method were utilized in previous MRP-supertree studies (e.g., Purvis, 1995a; Bininda-Emonds *et al.*, 1999; Daubin *et al.*, 2001; Liu *et al.*, 2001; Jones *et al.*, 2002; Salamin *et al.*, 2002), the result undoubtedly would have been very poorly resolved supertrees.

### 3.4 Taxonomic comprehensiveness

Proponents of supertrees have argued that comprehensive phylogenies that include all species in a group can be made using MRP, and that this is a potential advantage of the method and other supertree procedures (Purvis, 1995a; Bininda-Emonds *et al.*, 1999, 2002; Jones *et al.*, 2002). Before

presenting his MRP supertree of Primates, Purvis (1995a:405) noted that “It would not be feasible [presently] to compile a morphological or molecular data set representing all or even most of the living primate species.” Likewise, in their MRP study of placental mammals, Liu *et al.* (2001:1786) argued that a supermatrix analysis “remains presently unfeasible for the infraclass, because of several practical and analytical limitations . . .”

Given that Purvis (1995a) estimated the number of extant primate species at 203, that Liu *et al.* (2001) analyzed only 91 taxa in their mammalian study, and that several MP analyses of primary character data that included hundreds of taxa have been published in recent years (e.g., Van de Peer and de Wachter, 1997; Källersjö *et al.* 1998; Soltis *et al.*, 1999), analytical and computational differences between MRP-supertree analysis and supermatrix analysis are not relevant. A crucial limiting factor in large MRP or supermatrix parsimony searches is the number of taxa. Therefore, we do not think that a supermatrix study of 91 or 203 taxa would be any more difficult than a published MP analysis of character data from over 2500 taxa (Källersjö *et al.*, 1998; see also Section 3.3).

Many primate species, and most of the mammalian taxa analyzed in Liu *et al.* (2001), have been sampled already at the molecular level for mitochondrial genes, so there are no insurmountable stumbling blocks to building supermatrices for primates or families and orders of placental mammals. Total-evidence analyses of such groups are feasible in the age of genomics given a little initiative. This would require collecting some new data, but this is what systematists are paid to do. Construction of a supertree, given the published database, also is time consuming. As mentioned, data collection for the Liu *et al.* (2001) MRP analysis took over two years to complete. In those same two years, several genes could have been sequenced for the 91 taxa analyzed by these authors, and the sequence data could have been integrated with other published character data. Thus, arguments that supermatrix studies of different groups are not feasible presently are moot.

MRP supertrees with complete sampling of extant species have been published (e.g., Jones *et al.*, 2002), but these topologies were, in part, derived from informal classifications (see Sections 2.2 and 3.1). Again, we suggest that if no explicit character information for a particular taxon exists, the most useful response from a systematist would be to collect new character data for that taxon, not to include that taxon in a supertree analysis. In a review of supertrees, Bininda-Emonds *et al.* (2002) argued repeatedly that comprehensive phylogenetic hypotheses that included all members of particular groups have been crucial to the execution of several comparative evolutionary studies (e.g., Purvis *et al.*, 1995; Gittleman and Purvis, 1998), and that these studies would not have been possible without MRP-supertree analysis. In our opinion, this research program puts the cart before the horse.

Until a solid framework of relationships based on actual character data is presented, such broad evolutionary studies are premature. We contend that, given a few years of research effort, most of the studies cited by Bininda-Emonds *et al.* (2002) could have been executed within a supermatrix context.

### 3.5 Performance in simulations

Computer simulations have shown that, under certain circumstances, MRP-supertree analysis can approximate total-evidence supermatrix results and produce accurate phylogenetic trees (Bininda-Emonds and Sanderson, 2001; for simulation results of other supertree methods, see Burleigh *et al.*, 2004; Lapointe and Levasseur, 2004, Piaggio-Talice *et al.*, 2004). Several authors have cited the MRP simulations as a justification for previously published supertrees or for the general utility of the MRP method (Bininda-Emonds and Sanderson, 2001; Liu *et al.*, 2001; Bininda-Emonds *et al.*, 2002; Jones *et al.*, 2002). Although we acknowledge that these simulations showed that MRP-supertree analysis might produce accurate phylogenetic reconstructions in some circumstances, it is important to recognize the specific assumptions of the simulations, and to compare these assumptions with conditions in empirical supertree studies. If the assumptions do not match empirical data in any way, then it could be argued that the simulations have little relevance.

In fact, the conditions in the computer simulations differed markedly from those in all published MRP-supertree data sets. In the simulations of Bininda-Emonds and Sanderson (2001), all component data sets had the same number of characters, the rate of evolution for each data set on each branch was identical, all characters were multistate nucleotides, and a single model of evolution was utilized. None of these conditions were duplicated in published MRP-supertree data sets (e.g., Purvis, 1995a; Bininda-Emonds *et al.*, 1999; Liu *et al.*, 2001; Jones *et al.*, 2002; Kennedy and Page, 2002). More importantly, unlike published supertrees, the simulated data in Bininda-Emonds and Sanderson (2001) did not include data redundancies, unnecessary assumptions of monophyly, and other appeals to authority (see Springer and de Jong, 2001; Gatesy *et al.*, 2002). Until these inconsistencies between theoretical and actual supertree data sets are sorted out, the relevance of the simulations to empirical studies is questionable. Therefore, we disagree strongly with Bininda-Emonds and Sanderson's (2001:575) conclusion that, based on their simulation results, "... published supertrees should be judged and used with about the same degree of confidence that a total evidence tree might." We see no logical relationship between the

published MRP supertrees and the simulations presented by Bininda-Emonds and Sanderson (2001).

#### 4. Conclusion

Thus far, MRP has been the method of choice for building large supertrees. Through more careful and detailed analysis, several problems encountered in published MRP studies can be corrected. Duplications of data and inclusion of poorly justified source trees can be avoided with some extra work. Overall, however, we do not acknowledge any convincing arguments for the superiority of MRP-supertree methods relative to supermatrix analyses. In MRP-supertree analysis, hidden character support and conflicts within data sets are ignored, and novel clades can emerge that are 1) contradicted by all source trees in the MRP data set, and 2) not supported by combined supermatrix analysis of the component data sets. These facts demonstrate that there is a distortion of source trees and character evidence in MRP analysis (Rodrigo, 1993; 1996; Ronquist, 1996; Bininda-Emonds and Bryant, 1998; Slowinski and Page, 1999; Wilkinson *et al.*, 2001; Goloboff and Pol, 2002; Bryant, 2004). No proposed weighting scheme either corrects for these effects adequately or can be applied generally to all systematic data sets, both character and distance-based. Thus, the arguments that differential weighting of matrix elements in MRP analysis is beneficial and that MRP-supertree data sets are more inclusive than total-evidence supermatrix studies, are incompatible in practice. Unlike parsimony and likelihood methods that rely on direct analysis of character data, a logical justification for the MRP method is still lacking (Wilkinson *et al.*, 2001). Resolution of conflict among source trees is not explicable in terms of evolutionary events (Rodrigo, 1996; Slowinski and Page, 1999). Preliminary computer simulations do show that MRP-supertree analysis can mimic supermatrix results under unrealistic circumstances, but, to this point, the simulated data (e.g., of Bininda-Emonds and Sanderson, 2001) do not approximate published MRP data sets. Other possible advantages of MRP supertrees such as cost effectiveness, taxonomic comprehensiveness, and computational efficiency are debatable. We acknowledge that strict or semi-strict supertree methods (e.g., Steel, 1992; Goloboff and Pol, 2002) will be required ultimately to cobble together large supermatrix topologies into a comprehensive Tree of Life, but MRP should not be the method of choice for building this ultimate phylogeny.

## Acknowledgements

NSF grants to J. Gatesy (DEB-9985847, DEB-0213171, and DEB-0212572) and M. Springer (DEB-9903810) provided funding for this work. We thank B. Baum and M. Wilkinson for helpful comments, and O. Bininda-Emonds for inviting us to contribute a chapter to this volume.

## References

- BAKER, R. H. AND DESALLE, R. 1997. Multiple sources of character information and the phylogeny of Hawaiian drosophilids. *Systematic Biology* 46:654–673.
- BARRETT, M., DONOGHUE, M. J., AND SOBER, E. 1991. Against consensus. *Systematic Zoology* 40:486–493.
- BAUM, B. R. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon* 41:3–10.
- BININDA-EMONDS, O. R. P. 2003. Novel versus unsupported clades: assessing the qualitative support for clades in MRP supertrees. *Systematic Biology* 52:839–848.
- BININDA-EMONDS, O. R. P. AND BRYANT, H. N. 1998. Properties of matrix representation with parsimony analyses. *Systematic Biology* 47:497–508.
- BININDA-EMONDS, O. R. P., AND SANDERSON, M. J. 2001. Assessment of the accuracy of matrix representation with parsimony analysis supertree construction. *Systematic Biology* 50:565–579.
- BININDA-EMONDS, O. R. P., GITTLEMAN, J. L., AND PURVIS, A. 1999. Building large trees by combining phylogenetic information: a complete phylogeny of the extant Carnivora (Mammalia). *Biological Review* 74:143–175.
- BININDA-EMONDS, O. R. P., GITTLEMAN, J. L., AND STEEL, M. A. 2002. The (super)tree of life: procedures, problems, and prospects. *Annual Review of Ecology and Systematics* 33:265–289.
- BININDA-EMONDS, O. R. P., JONES, K. E., PRICE, S. A., CARDILLO, M., GRENYER, R., AND PURVIS, A. 2004. Garbage in, garbage out: data issues in supertree construction. In O. R. P. Bininda-Emonds (ed). *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 267–280. Kluwer Academic, Dordrecht, the Netherlands.
- BOOKSTEIN, F. L. 1994. Can biometrical shape be a homologous character? In B. Hall (ed.), *Homology: The Hierarchical Basis of Comparative Biology*, pp. 198–227, Academic Press, New York.
- BREMER, K. 1994. Branch support and tree stability. *Cladistics* 10:295–304.
- BRYANT, H. N. 2004. The cladistics of matrix representation with parsimony analysis. In O. R. P. Bininda-Emonds (ed). *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 353–368. Kluwer Academic, Dordrecht, the Netherlands.
- BURLEIGH, J. G., EULENSTEIN, O., FERNANDEZ-BACA, D., AND SANDERSON, M. J. 2004. MRF supertrees. In O. R. P. Bininda-Emonds (ed). *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 65–85. Kluwer Academic, Dordrecht, the Netherlands.
- COTTON, J. A. AND PAGE, R. D. M. 2004. Tangled trees from molecular markers: reconciling conflict between phylogenies to build molecular supertrees. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 107–125. Kluwer Academic, Dordrecht, the Netherlands.

- DAUBIN, V., GOUY, M., AND PERRIERE, G. 2001. Bacterial molecular phylogeny using supertree approach. *Genome Informatics* 12:155–164.
- FARRIS, J. S. 1983. The logical basis of phylogenetic analysis. In N. Platnick and V. Funk (eds.), *Advances in Cladistics*, volume 2, pp. 7–36, Columbia University Press, New York.
- FARRIS, J. S., KÄLLERSJÖ, M., AND DELAET, J. 2001. Branch lengths do not indicate support—even in maximum likelihood. *Cladistics* 17:298–299.
- FELSENSTEIN, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39:783–791.
- GATESY, J., MILINKOVITCH, M. C., WADDELL, V., AND STANHOPE, M. S. 1999. Stability of cladistic relationships between Cetacea and higher-level artiodactyl taxa. *Systematic Biology* 48:6–20.
- GATESY, J., MATTHEE, C., DESALLE, R., AND HAYASHI, C. Y. 2002. Resolution of a supertree / supermatrix paradox. *Systematic Biology* 51:652–664.
- GITTLEMAN, J. L. AND PURVIS, A. 1998. Body size and species richness in carnivores and primates. *Philosophical Transactions of the Royal Society of London B* 265:113–119.
- GOLOBOFF, P. A., AND POL, D. 2002. Semi-strict supertrees. *Cladistics* 18:514–525.
- HOLLAR, L. J. AND SPRINGER, M. S. 1997. Old World fruitbat phylogeny: evidence for convergent evolution and an endemic African clade. *Proceedings of the National Academy of Sciences of the United States of America* 94:5716–5721.
- HOOFER, S. R. AND VAN DEN BUSSCHE, R. A. 2001. Phylogenetic relationships of plecotine bats and allies based on mitochondrial ribosomal sequences. *Journal of Mammalogy* 82:131–137.
- HOOVER, A. 2001. A first: a (nearly) complete road map for the evolution of placental mammals. *University of Florida News, March 1*.
- JONES, K. E., PURVIS, A., MACLARNON, A., BININDA-EMMONDS, O. R. P., AND SIMMONS, N. B. 2002. A phylogenetic supertree of the bats (Mammalia: Chiroptera). *Biological Review* 77:223–259.
- KÄLLERSJÖ, M., FARRIS, J. S., CHASE, M. W., BREMER, B., FAY, M. F., HUMPHRIES, C. J., PETERSON, G., SEBERG, O., AND BREMER, K. 1998. Simultaneous parsimony jackknife analysis of 2538 *rbcL* DNA sequences reveals support for major clades of green plants, land plants, seed plants, and flowering plants. *Plant Systematics and Evolution* 213:259–287.
- KENNEDY, M. AND PAGE, R. D. M. 2002. Seabird supertrees: combining partial estimates of procellariiform phylogeny. *The Auk* 119:88–108.
- KLUGE, A. G. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). *Systematic Zoology* 38:7–25.
- LAPOINTE, F.-J. AND LEVASSEUR, C. 2004. Everything you always wanted to know about the average consensus, and more. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 87–105. Kluwer Academic, Dordrecht, the Netherlands.
- LEE, M. S. Y. AND HUGALL, A. F. 2003. Partitioned likelihood support and the evaluation of data set conflict. *Systematic Biology* 52:15–22.
- LIU, F.-G. R., MIYAMOTO, M. M., FREIRE, N. P., ONG, P. Q., TENNANT, M. R., YOUNG, T. S., AND GUGEL, K. F. 2001. Molecular and morphological supertrees for eutherian (placental) mammals. *Science* 291:1786–1789.
- MADDISON, W. P. 1997. Gene trees in species trees. *Systematic Biology* 46:523–536.
- MADSEN, O., SCALLY, M., DOUADY, C. J., KAO, D. J., DEBRY, R. W., ADKINS, R., AMRINE, H. M., STANHOPE, M. J., DE JONG, W. W., AND SPRINGER, M. S. 2001. Parallel adaptive radiations in two major clades of placental mammals. *Nature* 409:610–614.

- MARKS, J., SCHMID, C. W., AND SARICH, V. M. 1988. DNA hybridization as a guide to phylogeny: relations of the Hominoidea. *Journal of Human Evolution* 17:769–786.
- MIYAMOTO, M. M. 1985. Consensus cladograms and general classifications. *Cladistics* 1:186–189.
- MIYAMOTO, M. M. AND FITCH, W. M. 1995. Testing species phylogenies and phylogenetic methods with congruence. *Systematic Biology* 44:64–76.
- MURPHY, W. J., EIZRIK, E., JOHNSON, W. E., ZHANG, Y. P., RYDER, O. A. AND O'BRIEN, S. J. 2001. Molecular phylogenetics and the origins of placental mammals. *Nature* 409:614–618.
- NOVACEK, M. J. 1980. Phylogenetic analysis of the chiropteran auditory region. In D. Wilson and A. Gardner (eds), *Proceedings of the Fifth International Bat Research Conference*, pp. 317–330. Texas Tech. University, USA.
- PATTERSON, C. 1982. Morphological characters and homology. In A. Joysey and A. Friday (eds), *Problems of Phylogenetic Reconstruction*, pp. 21–74. Academic Press, London.
- PIAGGIO-TALICE, R., BURLEIGH, J. G., AND EULENSTEIN, O. 2004. Quartet supertrees. In O. R. P. Bininda-Emonds (ed). *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 173–191. Kluwer Academic, Dordrecht, the Netherlands.
- PISANI, D., YATES, A., LANGER, M., AND BENTON, M. 2001. A genus-level supertree of the Dinosauria. *Proceedings of the Royal Society of London Series B* 269:915–921.
- PISANI, D. AND WILKINSON, M. 2002. Matrix representation with parsimony, taxonomic congruence, and total evidence. *Systematic Biology* 51:151–155.
- PURVIS, A. 1995a. A composite estimate of primate phylogeny. *Philosophical Transactions of the Royal Society of London B* 348:405–421.
- PURVIS, A. 1995b. A modification to Baum and Ragan's method for combining phylogenetic trees. *Systematic Biology* 44:251–255.
- PURVIS, A., NEE, S., AND HARVEY, P. H. 1995. Macroevolutionary inferences from primate phylogeny. *Proceedings of the Royal Society of London B* 260:329–333.
- RAGAN, M. A. 1992. Phylogenetic inference based on matrix representation of trees. *Molecular Phylogenetics and Evolution* 1:53–58.
- DE QUEIROZ, K. AND POE, S. 2001. Philosophy and phylogenetic inference: a comparison of likelihood and parsimony methods in the context of Karl Popper's writings on corroboration. *Systematic Biology* 50:305–321.
- RODRIGO, A. G. 1993. A comment on Baum's method for combining phylogenetic trees. *Taxon* 42:631–636.
- RODRIGO, A. G. 1996. On combining cladograms. *Taxon* 45:267–274.
- RONQUIST, F. 1996. Matrix representation of trees, redundancy, and weighting. *Systematic Biology* 45:247–253.
- SALAMIN, N., HODKINSON, T. R., AND SAVOLAINEN, V. 2002. Building supertrees: an empirical assessment using the grass family (Poaceae). *Systematic Biology* 51:136–150.
- SANDERSON, M. J., PURVIS, A., AND HENZE, C. 1998. Phylogenetic supertrees: assembling the trees of life. *Trends in Ecology and Evolution* 13:105–109.
- SEMPLE, C. AND STEEL, M. 2000. A supertree method for rooted trees. *Discrete Applied Mathematics* 105:147–158.
- SIBLEY, C. G. AND AHLQUIST, J. E. 1990. *Phylogeny and Classification of Birds: a Study in Molecular Evolution*. Yale University Press, New Haven.
- SIMMONS, N. B. AND GEISLER, J. H. 1998. Phylogenetic relationships of *Icaronycteris*, *Archaeonycteris*, *Hassianycteris*, and *Palaeochiropteryx* to extant bat lineages, with comments on the evolution of echolocation and foraging strategies in Microchiroptera. *Bulletin of the American Museum of Natural History* 235:1–82.

- SLOWINSKI, J. B. AND PAGE, R. D. M. 1999. How should species phylogenies be inferred from sequence data? *Systematic Biology* 48:814–825.
- SOLTIS, P. S., SOLTIS, D. E., AND CHASE, M. W. 1999. Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. *Nature* 402:402–404.
- SPRINGER, M. S. AND DE JONG, W. W. 2001. Phylogenetics. Which mammalian supertree to bark up? *Science* 291:1709–1711.
- STEEL, M. A. 1992. The complexity of reconstructing trees from qualitative characters and subtrees. *Journal of Classification* 9:91–116.
- SWIDERSKI, D. L., ZELDITCH, M. L., AND FINK, W. L. 1998. Why morphometrics is not special: coding quantitative data for phylogenetic analysis. *Systematic Biology* 47:508–519.
- SWOFFORD, D. L. 1998. *PAUP\*. Phylogenetic Analysis Using Parsimony (\*and Other Methods)*. Version 4. Sinauer, Sunderland, Massachusetts.
- TEELING, E. C., SCALLY, M., KAO, D. J., ROMAGNOLI, M. L., SPRINGER, M. S., AND STANHOPE, M. J. 2000. Molecular evidence regarding the origin of echolocation and flight in bats. *Nature* 403:188–192.
- VAN DEN BUSSCHE, R. A. AND HOOFER, S. R. 2001. Evaluating monophyly of Nataloidea (Chiroptera) with mitochondrial DNA sequences. *Journal of Mammalogy* 82:320–327.
- VAN DE PEER, Y. AND DE WACHTER, R. 1997. Evolutionary relationships among the eukaryotic crown taxa taking into account site-to-site rate variation in 18S rRNA. *Journal of Molecular Evolution* 45:619–630.
- WILKINSON, M., THORLEY, J. L., LITTLEWOOD, D. T. J., AND BRAY, R. A. 2001. Towards a phylogenetic supertree of Platyhelminthes? In D. Littlewood and R. Bray (eds), *Interrelationships of the Platyhelminthes*, pp. 292–301. Chapman-Hall, London.
- WILKINSON, M., LAPONTE, F.-J., AND GOWER, D. J. 2003. Branch lengths and support. *Systematic Biology* 52:127–130.
- WOJCIECHOWSKI, M. F., SANDERSON, M. J., STEEL, K. P., AND LISTON, A. 2000. Molecular phylogeny of the “temperate herbaceous tribes” of papilionoid legumes: a supertree approach. In P. Herendeen and A. Bruneau (eds), *Advances in Legume Systematics* 9:277–298. Royal Botanic Garden, Kew.

## **5. Supertrees and their applications**

## Chapter 18

# SUPERTREES, COMPONENTS AND THREE-ITEM DATA

David M. Williams

**Abstract:** Supertree construction is explored from the perspective of binary and three-item data. Binary data (components) code groups and subgroups, three-item data code relationships. Data are corroborative, consistent or conflicting. For either components or three-item statements, corroborative data support the same group; consistent data support different, non-conflicting groups; and conflicting data suggest alternative solutions. Binary data, when analyzed using parsimony, are unable to resolve simple cases of conflict. The lack of resolution is a result of the nature of the data *and* the peculiarities of “optimization”. Three-item data resolve most cases of conflict. Problems in systematic data analysis might be improved by investigations relating to the data, their meaning and their representation rather than exploring more methods.

**Keywords:** binary data; components; conflict; parsimony; three-item data

### 1. Introduction

In the past few years, “supertrees” have become popular (although see Gordon, 1986; Page, 1994:64), described by Sanderson *et al.* (1998:105) as “Any such tree containing all the taxa from a collection of trees...” and a “strict supertree” (or “consistent supertree”; Wilkinson *et al.*, 2001) as “one that agrees with all the trees from which it was derived”. A supertree can be viewed as a kind of consensus tree (Bininda-Emonds *et al.*, 2002:267), the principle difference being that consensus tree techniques require taxa (terminals) in each fundamental cladogram (*sensu* Nelson, 1979) to be identical, whereas supertrees do not. Thus, supertrees include overlapping

fundamental cladograms, so that they can deal with more general problems, such as finding solutions for cladogram combinations like A(B(CD)) and B(C(DE)), and A(B(CD)) and C(DE(FG)).

Because most of the usual consensus techniques require fundamental cladograms to have the same terminal taxa, Baum and Ragan suggested independently an alternative method based on representing the data in a matrix format (Baum, 1992; Ragan, 1992a, b; see also Doyle, 1992). For every fundamental cladogram, each node can be represented in a data matrix by using the additive binary coding procedure of Farris *et al.* (1970; see also Farris, 1973). The resulting matrix can then be analyzed using a parsimony program to find a (or some) summary solution that represents the supertree (Baum, 1992; Ragan, 1992a, b; Baum and Ragan, 1993, 2004; see also Rodrigo, 1993, 1996; Williams, 1994, 1996a). The conversion of fundamental cladograms to matrix entries removed the requirement for all terminals to be represented in each cladogram. The approach, named “matrix representation with parsimony” (MRP; Ragan, 1992a), has engendered considerable discussion, especially because some consider supertrees to be useful summaries of many different data sets (e.g., see the reviews of Wilkinson *et al.*, 2001; Bininda-Emonds *et al.*, 2002).

The focus of this paper is the use of matrix representation methods for encoding tree topologies and the implementation of parsimony for analyzing these representations. In so doing, it is necessary to address two interconnected questions: 1) the use of binary data to represent cladograms and 2) the limitations of parsimony as implemented in current computer programs. For the latter, the significant feature of current versions of parsimony is optimization, the method behind finding the best fitting cladogram relative to the data.

## 2. Data

Before discussing various consensus methods, Nixon and Carpenter (1996:their table 1) distinguished certain kinds of trees, which they divided into “direct” (data-derived trees) and “indirect” (summary trees derived from trees). “Direct” trees are those found from the analysis of binary data matrices of original character information using methods such as parsimony, compatibility, or maximum likelihood. They subdivided “indirect” trees further into “consensus” and “compromise” trees, the former including only strict trees (or, in Carpenter and Nixon’s terms, “Nelson” consensus trees), whereas the latter included most other commonly used consensus techniques, such as Adams, combinable component, majority-rule, and “other component approaches” (which would include MRP; Carpenter, pers.

comm.) including three-taxon statements (Nixon and Carpenter, 1996:307, their table 1). Coincidentally, Wilkinson *et al.* (2001; see also Wilkinson and Thorley, 1998) divided supertree methods into “direct” and “indirect”, with the former referring to all (or nearly all) the usual consensus techniques, and the latter referring to matrix methods, such as MRP or three-item statements. The contrast between Nixon and Carpenter and Wilkinson *et al.* reflects differing points of view concerning the issue of basic (“raw”) data and how they might be represented relative to any particular problem. For character data, binary representation is seen as primary (Nixon and Carpenter, 1996; see also Bininda-Emonds and Bryant, 1998:498); for consensus trees and supertrees, cladograms are seen as primary (Wilkinson *et al.*, 2001).

Here, I note that both Nixon and Carpenter and Wilkinson *et al.* consider three-item data to be “indirect”. That is, three-item statements are understood to be “implicitly tree-derived” (Nixon and Carpenter, 1996:307, their table 1; Wilkinson *et al.* 2001) because suites of three-item statements are derived from binary variables or fundamental cladograms under both systems. The distinction is artificial because it is based on two false assumptions. The first assumption is that there is an agreed and accepted way to represent primary data; the second assumption is that there is no direct relationship between a matrix entry and a cladogram. There is no reason to assume that three-item data cannot be viewed as “direct” inasmuch as all data, binary or otherwise, are simply a representation. More significantly, it is evident that three-item data can be represented either as a cladogram, A(BC), or as a matrix entry, 0(11). Rather, a better distinction is between binary and three-item data. Binary data represent *groups* (or subgroups, Wilkinson, 1994a), whereas three-item data represent *relationships* (Nelson and Platnick, 1991; Nelson and Ladiges, 1994). This distinction is expanded on below.

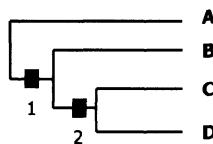
## 2.1 Groups and component coding

For the cladogram A(B(CD)) (Figure 1), there are two components. Component 1 represents the group B + C + D (BCD) and component 2 represents the group C + D (CD). Thus,

$$\text{Component} = \text{Group}.$$

Each component is equivalent to a node on a cladogram (Nelson, 1979:3; Bininda-Emonds and Bryant, 1998:499; see Figure 1). Thus,

$$\text{Component} = \text{Group} = \text{Node}.$$



*Figure 1.* Cladogram representing the interrelationships among four taxa A–D, with data supporting nodes (components) 1 and 2 (components marked with shaded boxes).

*Table 1.* Matrix representation of the four taxa (rows A–D) in Figure 1 by two “characters” (columns 1–2).

|   | 1 | 2 |
|---|---|---|
| A | 0 | 0 |
| B | 1 | 0 |
| C | 1 | 1 |
| D | 1 | 1 |

The groups specified by the nodes are equivalent to taxonomic groups, in the sense that B + C + D and C + D can (but need not) be named taxa. Thus,

$$\text{Component} = \text{Group} = \text{Node} = \text{Taxon}.$$

The cladogram in Figure 1 can also be summarized as a matrix of two binary “characters” (Table 1). The matrix in Table 1 has four taxa (rows A–D) and two “characters” (columns 1–2). The “characters” in the matrix and the components of the cladogram match exactly (Figure 1). The argument has been extended such that “shared characters” can be considered equivalent to components (Page, 1987:6); therefore,

$$\text{Component} = \text{Group} = \text{Node} = \text{Taxon} = \text{Shared Character}.$$

Components can be understood also as “general statements of synapomorphy” (Nelson and Platnick, 1981:169; see also Nelson, 1979:8) or general statements of “homology” inasmuch as the component (homology) specifies a particular group. Therefore,

$$\text{Component} = \text{Group} = \text{Node} = \text{Taxon} = \text{Shared Character} = \text{Homology}.$$

This is the usual way of looking at systematic data, at least from the perspective made popular by numerical taxonomy (Sokal and Sneath, 1963;

see Platnick, 1993:271): components (and reduced portions of them, Wilkinson, 1994a) are equivalent to groups.

Combining components notes simply those groups that the individual components share. Thus, A(BCD) and AB(CD) sum to A(B(CD)). There is no reason for components to contain identical taxa; therefore, they might be appropriate for supertree construction. For example, A(B(CD)) and B(C(DE)) sum to A(B(C(DE))). No method is required to arrive at this conclusion. It might be suggested that these kinds of problems are trivial, in the sense that they yield “strict” or “consistent” supertrees. Only conflicting components present problems as certain combinations of components conflict in a seemingly absolute sense. Consider the simple example AB(CD) + AC(BD). When the groups CD and BD are the units of combination, they conflict because there is no possible solution that could include both groups.

If “binary character” is substituted for “shared character” (cf., Page, 1987) and homologue is substituted for homology (Nelson, 1989, 1994; Williams and Humphries, 2003b; Williams, in press), then the equation simply denotes one aspect of the data:

$$\text{Component} = \text{Group} = \text{Node} = \text{Binary character (s)} = \text{Homologue}.$$

These data are really equivalent to the phenetic notion of similarity, even if some of the terms are intended to imply group-membership. More significantly, these kinds of data can be rendered informative only by the application of a method. To rely on a method for sorting groups suggests that solutions are imposed rather than discovered, and the imposition is only as good (or as effective) as the method. Nevertheless, such methods fail to deal with simple conflicting components such as AB(CD) + AC(BD).

One possible solution is to include both groups in the same summary, yielding the solution A(CD)(BD). Such a solution contains redundancy, inasmuch as taxon D occurs twice. One occurrence of taxon D can be removed from each solution to yield A(CD)(B) and A(C)(BD), which are summarized effectively as A(BCD).

## 2.2 Relationships and three-item coding

The problem of conflict resides with the idea that *groups* are the basic unit for analysis (Nelson, 1989:277). By itself, any group, such as ABC, is uninformative of relationships. Likewise, each of its included subgroups, such as AB, is uninformative of relationships. Taken together (ABC + BC), the data are rendered informative when the relationship is specified accurately: A(BC) (Nelson and Ladiges, 1991:481). Of the items listed in the previous section, component, node, and taxon appear equivalent to *group* (a

*Table 2.* Three-item data for the two components AB(CD) and AC(BD) and their solution.

|       |   |       |   |       |   |          |
|-------|---|-------|---|-------|---|----------|
| A(CD) | + | B(CD) | + | A(BD) | = | A(B(CD)) |
| A(CD) | + | C(BD) | + | A(BD) | = | A(C(BD)) |

taxon is conceived usually of as a named group of organisms). In contrast to component coding, three-item coding sees each node as a *relation* between branches (Nelson and Ladiges, 1996). Three-item coding relates some branches more closely than other branches of the cladogram, with each separate relation expressed as a three-item statement: for example, A(BC), where B and C are related to each other more closely than either are to A. When data are dealt with as specific, minimal statements of relationship, the above equation might be improved further (see Nelson, 1989, 1994):

$$\text{Relationship} = \text{Taxon} = \text{Homology}.$$

The two conflicting components from before, AB(CD) and AC(BD), can be represented by two three-item statements each: A(CD) and B(CD), and A(BD) and C(BD), respectively. These statements can be summed to arrive at a solution that includes as many statements as possible. In this case, there are two solutions, each including three statements: A(B(CD)) and A(C(BD)) (Table 2). All the statement can be summarized by the cladogram A(BCD) (Nelson, 1996). Thus, data are rendered informative by being associated with a cladogram directly and a summary of those relationships are discovered rather than imposed (via a method).

One could, of course, simply reduce the problem of conflict to a lack of data, but that handicap betrays the very necessity of supertrees as anything more than interim and unstable summaries. As Wilkinson *et al.* (2001:300) note, “In the absence of additional data, resolution of conflict can be justified by explanation and understanding of the conflict”. Explanation to one side, the understanding resides in use of groups.

### 3. Methods

#### 3.1 The dilemma of conflict

“Isms multiply when ideologies collide. Strange though it may seem at “the end of history,” words we smile or scoff at were once *casus belli*, fought over like territory, flung about like grenades...Coined in

innocence or forged in anger, words like these became “calls to battle”...When their “logical meaning” was spent, they still retained a “magical power” to provoke or persuade “simply by being used” (Moore, 2001).

Probably the first attempt to acknowledge the issue of conflicting data in a numerical context was by Sokal and Sneath (1963:225–226; see also Sneath, 1988:266, 1995:287), who explored what became known as “Hennig’s dilemma” (Felsenstein, 1982; Felsenstein, 1984:170), or, as Farris and Kluge (1997:216) preferred, “Felsenstein’s dilemma”. Felsenstein (1982) explored “Hennig’s dilemma” (Felsenstein, 1982:380, 1984:his table 10.1), but did more than simply restate the problem because he suggested two analytical methods by which character conflict might be resolved: “parsimony” and “compatibility”. He characterized “parsimony” as

“mak[ing] a compromise among the characters, a compromise with which no individual character may be entirely compatible”,

and “compatibility” as

“find[ing] the phylogeny that is completely compatible with a plurality of characters, even though the remaining characters may be extremely incompatible with it.” (Felsenstein, 1982:381, 1984:172).

This distinction was made a few years earlier by Farris and Kluge:

“The distinction is that compatibility methods recognize only perfect correlations — sets of fully congruent characters — whereas the Wagner [parsimony] method more realistically accepts some imperfect correlations, which makes possible a better fit to all available evidence.” (Farris and Kluge, 1979:405; Kluge, 1984:28).

Compatibility methods (Nelson, 1979; Page, 1989; see also Ross and Rodrigo, 2004) and parsimony methods (as MRP; Baum, 1992; Ragan, 1992a, b; Baum and Ragan, 1993, 2004) have both been implemented to solve consensus and supertree problems.

The contrast has always been understood as distinguishing between two different kinds of analysis. An alternative view is that the methods highlight how data might be viewed differently, with compatibility using only “whole” characters, whereas parsimony can use “partial” characters (i.e., the “imperfect correlations”). Thus, both parsimony and compatibility methods can be understood as parsimony methods in the sense that they minimize (or maximize) a particular quantity (Felsenstein, 1982). In the same way, analysis of three-item data is a parsimony analysis because it maximizes the number of three-item statements included in the final, summary cladogram.

At present, three-item analysis is implemented using the suite of programs found in the TAX package (published by G. J. Nelson and P. Y. Ladiges), which enables the automatic conversion of any standard binary matrix into its three-item equivalent. The three-item matrix can then be analyzed using any current parsimony program. This is not necessarily the best way of discovering the optimal tree (for further comment, see Williams and Siebert, 2000). A new program, 3item (published by M. C. Ebach *et al.*), deals directly with the data (the statements) rather than using matrices and optimization procedures. Nevertheless, when using the three-item matrix approach, parsimony programs will work efficiently enough in spite of the technical procedures inherent in their execution because each statement fits a summary cladogram exactly (one step) or is rejected (two steps). Because statements fit cladograms with either one or two steps, the approach is somewhat similar to clique (compatibility) methods (Wilkinson, 1994b; but see comments in Williams and Siebert, 2000 and below). Regardless of implementation, it is the representation of the data that appears to be crucial.

## 4. Analysis of some simple supertree problems

### 4.1 Parsimonious analysis (MRP) of binary data

“His promises were, as he then was,  
mighty;  
But his performance, as he is now,  
nothing”

(Henry VIII, Act 4, Scene 2)

To understand the mechanics behind MRP is to understand the workings of parsimony as it is implemented currently. This is best appreciated with a few simple examples. For instance, parsimony analysis of the two components DE(ABC) and CE(ABD) finds four equally parsimonious, unique solutions (Table 3). Consider solution 1: DE(C(AB)). Component DE(ABC) supports the basal node (1 step), whereas component CE(ABD) supports the terminal node, but only partially, with A + B recognized as a group (1 step for AB, 1 step for C). Consider solution 2: CE(D(AB)). Component CE(ABD) supports the basal node (1 step), whereas component DE(ABC) supports the terminal node, but only partially, with A + B recognized as a group (1 step for AB, 1 step for C). Consider solution 3: E(D(ABC)). Component DE(ABC) supports the terminal node (1 step), whereas component CE(ABD) supports the basal node, but only partially, with the A, B, and D

*Table 3.* Four equally parsimonious cladograms from parsimony analysis (MRP) of DE(ABC) + CE(ABD). Parsimony analysis yields six cladograms, two of which (not shown) include one node supported only by a zero-length branch. For these two cladograms, E(D(C(AB))) reduces to E(D(ABC)) (cladogram 3), E(C(D(AB))) reduces to E(C(ABD)) (cladogram 4).

| <b>DE(ABC) + CE(ABD)</b> |
|--------------------------|
| 1. DE(C(AB))             |
| 2. CE(D(AB))             |
| 3. E(D(ABC))             |
| 4. E(C(ABD))             |

supporting the ABCD group (1 extra step for the “reversal” in C). Finally, consider solution 4: E(C(ABD)). Component CE(ABD) supports the terminal group (1 step), whereas DE(ABC) supports the basal node E(ABCD), but only partially, with A, B and C supporting the ABCD group (1 extra step for the “reversal” in D). Thus, parsimony modifies the component data to “fit” different cladograms by creating new groups and subgroups; it accepts “some imperfect correlations, which makes possible a better fit to all available evidence” (Farris and Kluge, 1979:405). Regardless of the various solutions, the overall result (the strict consensus of all four cladograms) is a bush, an uninformative solution.

A direct inspection of the two original components reveals that there is one common group, A + B, suggesting that CDE(AB) might be a reasonable solution regardless of the workings of parsimony (it is the strict consensus of solutions 1 and 2 in Table 3). In addition, component DE(ABC) suggests that C is related more closely to A and B than either is to E or D, and component CE(ABD) suggests that D is related more closely to A and B than either is to E or C. Together, this suggests further that E(CD(AB)) is a better summary than CDE(AB). No method is required to arrive at this conclusion beyond examination of the evidence. Parsimony complicates matters by its method of counting to form groups and subgroups; hence, all four cladograms are considered equally useful summaries (in the sense that they are equally parsimonious) and the overall solution (the strict consensus) is inconclusive.

In supertree problems, one or more taxa might be missing. If taxon E is missing from component DE(ABC), parsimony analysis still finds no informative solution to D(ABC) + CE(ABD), in spite of the AB group still being common to both components, and C and D being related more closely to A and B than they are to E. Various other permutations with missing taxa perform equally badly in the sense that common groups and subgroups seem to be almost impossible to discover (Table 4).

*Table 4.* Eight two-component problems and their results. Seven of the combinations lack one taxon and thus fall into the supertree setting. Although parsimony will find a suite of solutions in most cases, the solutions conflict to the extent that there is no summary cladogram (i.e., no informative strict consensus = “Conflict”).

| <b>DE(ABC) +</b><br><b>CE(ABD)</b> | <b>D(ABC) +</b><br><b>CE(ABD)</b> | <b>E(ABC) +</b><br><b>CE(ABD)</b> | <b>DE(ABC) +</b><br><b>E(ABD)</b> |
|------------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|
| Conflict                           | Conflict                          | E(C(ABD))                         | E(D(ABC))                         |
| <b>DE(ABC) +</b><br><b>C(ABD)</b>  | <b>DE(AB) +</b><br><b>CE(ABD)</b> | <b>DE(BC) +</b><br><b>CE(ABD)</b> | <b>DE(AC) +</b><br><b>CE(ABD)</b> |
| Conflict                           | CE(D(AB))                         | Conflict                          | Conflict                          |

## 4.2 Parsimonious analysis of three-item data

Parsimony analysis of three-item data for the two components DE(ABC) and CE(ABD) finds two solutions, E(D(C(AB))) and E(C(D(AB))), with their strict consensus providing the overall summary solution of E(CD(AB)) (Table 5).

DE(ABC) provides six three-item statements and CE(ABD) provides another six. The statement E(AB) is present in both cladograms; hence, there are a total of 12 statements, 11 of which are unique (Table 6).

The solution E(CD(AB)) summarizes eight three-item statements, seven of which are included in the data (Table 6, included statements in bold; E(AB) appears in both components; hence, a total of eight) and one, E(CD), is implied by the solution; four statements from the data are rejected. Three-item parsimony is not complicated by any unnecessary optimization procedure; hence, the data can be simply added together to find a summary cladogram. In addition, data are considered either true for the cladogram and are included or false and rejected. In this sense, one is able to examine the source of all particular rejected statements and investigate the reasons for their exclusion.

In the more usual supertree problems, where one or more taxa are missing, exact solutions are still possible. If taxon E is missing from the first component, the total number of three-item statements for it is reduced to three. For D(ABC) and CE(ABD), there are a total of nine three-item statements (Table 7), which result in two most parsimonious solutions, CE(D(AB)) and E(D(C(AB))).

The first cladogram, CE(D(AB)), implies a total of seven statements, all of which are found in the data; two statements, D(BC) and D(AC), are excluded. The second cladogram, E(D(C(AB))), implies a total of ten statements, seven of which are found in the data; two statements, C(AD) and

*Table 5.* Two equally parsimonious cladograms for the three-item data from DE(ABC) + CE(ABD). The strict consensus yields the solution E(CD(AB)).

| <b>DE(ABC) + CE(ABD)</b> |
|--------------------------|
| E(D(ABC))                |
| E(C(ABD))                |

*Table 6.* Three-item statement data for the two components DE(ABC) and CE(ABD).

| <b>DE(ABC)</b> | <b>CE(ABD)</b> |
|----------------|----------------|
| D(AB)          | C(AB)          |
| D(BC)          | C(AD)          |
| D(AC)          | C(BD)          |
| E(AB)          | E(AB)          |
| E(AC)          | E(AD)          |
| E(BC)          | E(BD)          |

*Table 7.* Three-item statement data for the two components D(ABC) and CE(ABD).

| <b>D(ABC)</b> | <b>CE(ABD)</b> |
|---------------|----------------|
| D(AB)         | C(AB)          |
| D(BC)         | C(AD)          |
| D(AC)         | C(BD)          |
| E(AB)         |                |
| E(AD)         |                |
| E(BD)         |                |

C(BD), are excluded (Table 8) and one, E(CD), is implied. Of the two solutions, the first, CE(D(AB)), seems to be the most efficient in terms of included statements derived directly from the data.

The same permutations with missing taxa that were explored in Table 4 all yield informative summary solutions (Table 9).

Three-item data can deal with conflict in a more efficient way primarily because groups and their subgroups are not considered the basic unit of analysis and optimization becomes irrelevant. Simple investigation of non-conflicting data indicates how further resolution might be achieved and how more nodes might be discovered (Nelson, 1996). From the perspective of supertrees and the gradual accumulation of relevant data, this is of some significance because data sets rarely include all taxa relevant to a particular problem. Suppose only three statements of relationship are known for a suite of five taxa, A–E: A(BC), B(CD) and C(DE). No computer program is required to conclude that the summary solution is A(B(C(DE))). The cladogram implies ten statements in total. If any of the seven statements not

*Table 8.* Two equally parsimonious cladograms for the three-item data from D(ABC) + CE(ABD).

| <b>D(ABC) + CE(ABD)</b> |              |              |              |
|-------------------------|--------------|--------------|--------------|
| <b>CE(D(AB))</b>        | <b>C(AB)</b> | <b>D(AB)</b> | <b>C(AB)</b> |
| <b>D(AB)</b>            | <b>C(AB)</b> | <b>D(AB)</b> | <b>C(AB)</b> |
| <b>D(BC)</b>            | <b>C(AD)</b> | <b>D(BC)</b> | <b>C(AD)</b> |
| <b>D(AC)</b>            | <b>C(BD)</b> | <b>D(AC)</b> | <b>C(BD)</b> |
|                         | <b>E(AB)</b> |              | <b>E(AB)</b> |
|                         | <b>E(AD)</b> |              | <b>E(AC)</b> |
|                         | <b>E(BD)</b> |              | <b>E(BC)</b> |

*Table 9.* Eight two-component problems and their results. Seven of the combinations lack one taxon.

|                                    |                                   |                                   |                                   |
|------------------------------------|-----------------------------------|-----------------------------------|-----------------------------------|
| <b>DE(ABC) +</b><br><b>CE(ABD)</b> | <b>D(ABC) +</b><br><b>CE(ABD)</b> | <b>E(ABC) +</b><br><b>CE(ABD)</b> | <b>DE(ABC) +</b><br><b>E(ABD)</b> |
| <b>E(CD(AB))</b>                   | <b>CE(D(AB))</b>                  | <b>E(C(ABD))</b>                  | <b>E(D(ABC))</b>                  |
| <b>DE(ABC) +</b><br><b>C(ABD)</b>  | <b>DE(AB) +</b><br><b>CE(ABD)</b> | <b>DE(BC) +</b><br><b>CE(ABD)</b> | <b>DE(AC) +</b><br><b>CE(ABD)</b> |
| <b>DE(C(AB))</b>                   | <b>CE(D(AB))</b>                  | <b>E(C(ABD))</b>                  | <b>E(C(ABD))</b>                  |

included in the original information are discovered subsequently, they will be consistent with the same solution. If any of A(BC), B(CD), or C(DE) is discovered subsequently, then they simply corroborate the existing solution. One might enumerate all the statements that can potentially conflict. In each case, the conflict would be with one specified node. If D(CE) were found, for example, the solution would simply reduce to A(B(CDE)), where the terminal trichotomy suggests three possible solutions: B(C(DE)), B(D(CE)), and B(E(CD)). In short, three-item data are more sensitive to how information accumulates to produce overall summary solutions, especially when all terminals are not represented.

### 4.3 Is there a general method of analysis?

If methods are seen as a source of imposition, is it possible to discover a generalized approach, one that is applicable to all kinds of data that are intended to discover relationships? The early argument between compatibility and parsimony seems to have been based partly on considerations of the data themselves, such that the former uses “whole” groups (components), whereas the latter allows for subgroups (partial components). The matter is more complicated because optimization, the

“factor” that allows parsimony to work efficiently, creates artifactual groupings out of the myriad of potential subgroups (e.g., Goloboff and Pol, 2002:their figure 1). Thus parsimony, as implemented in current computer programs, should be rejected as a general method of data analysis.

Nelson and Ladiges (2001:393) suggested recently that geographical data are corroborative, consistent, or conflicting. For either components or three-item statements, corroborative data support the same group; consistent data support different, non-conflicting groups; and conflicting data suggest alternative solutions. These three options suggest that there is a general method for reconciling systematic data and might help explain why this idea has occurred so frequently (Estabrook, 1972, Nelson, 1979; Morse and White, 1979; Nelson and Platnick, 1980; Page, 1987, 1990a, b; Patterson, 1982, 1988; Scotland, 1992:7, his figure 1.4, 1997, 2000:159; Lorenzen, 1993; Wägele, 1994; Kitching *et al.*, 1998:9, their figure 1.6; see commentary in Wilkinson, 1994c).

It might be premature to conclude that three-item data solutions are method independent, but results so far suggest that this is the case (Williams and Siebert, 2000; Nelson *et al.* 2003; Williams and Humphries, 2003a; Nelson, in press). By method independent, I mean that the data alone provide the answer, and as a means of discovery rather than of imposition. It seems data and their representation are the most significant factors. And, for the moment, it seems that three-item representation captures that relevant information the most efficiently.

## 5. Total evidence (“simultaneous analysis”) revisited

“As we see the issue, it is not consensus versus ‘combined data sets’ but rather the *nature of data* and how they are combined” (Nelson and Ladiges, 1991:481; emphasis added).

If three-item data provide a more useful approach to resolving conflict, the thorny issue of how useful, or significant, supertrees might be, relative to the analysis of pooled “character” data, remains (Kluge, 1989, 1998; Miyamoto, 1985; Miyamoto and Fitch, 1995). Barrett *et al.* (1991; reproduced in Eernisse and Kluge, 1993:figure 3 and Kluge and Wolf, 1993:figure 3) offered a simple example to demonstrate some differences between “character congruence” and “taxonomic congruence”. They presented two matrices, each with seven binary characters. Parsimony analysis of their matrix 1 found one cladogram, A(B(CD)); parsimony analysis of their matrix 2 found a different cladogram, A(D(BC)). The strict consensus of the two is A(BCD). Parsimony analysis of the pooled binary data (matrix 1 plus

matrix 2) found the unique cladogram (AC)(BD). Nelson (1993) applied parsimony analysis to the three-item data derived from both matrices 1 and 2 separately, finding the same two cladograms as the binary matrices. Parsimony analysis of the pooled three-item data also found the same two cladograms, with the strict consensus of A(BCD). Pisani and Wilkinson (2002) applied MRP to the two original cladograms and found both after parsimony analysis of that matrix, with A(BCD) as their strict consensus. Finally, application of three-item consensus to the two original cladograms also finds both, with A(BCD) as their strict consensus.

These analyses demonstrate first that MRP and “total evidence” might not arrive at the same result (as noted by Pisani and Wilkinson, 2002:152). Second, the analyses demonstrate that MRP, parsimony analysis of the pooled three-item data, and three-item consensus all arrive at the same solution (not noted by Pisani and Wilkinson, 2002). One might still ask which of the two solutions, (AC)(BD) or A(BCD), is the “best” or, perhaps, the most appropriate. It might seem that contrary to received wisdom (Barrett *et al.*, 1993), the anomalous result comes from the parsimony analysis of the pooled binary data such that the (AC)(BD) solution is an “optimization” artifact (Nelson, 1993): eight of the 14 characters require alteration to fit the optimal cladogram.

Chippindale and Wiens (1994) presented a similar example using two matrices with five taxa (A–E). The general results were similar to those of Barrett *et al.* Parsimony analyses of their matrices 1 and 2 found the cladograms C((AB)(DE)) and D((AB)(CE)), respectively, with the strict consensus of the two being CDE(AB). The pooled binary character data found instead the unique cladogram A(B(C(DE))). Parsimony analysis of the three-item data for the two matrices finds the same two cladograms as the binary data; parsimony analysis of the pooled three-item data finds a different unique cladogram, (AB)((C(DE))), which includes the AB node. MRP and three-item consensus find the two original cladograms with the strict consensus of CDE(AB).

One might be tempted to assume that the parsimony analysis of the pooled binary data is better in this case than the one-node consensus because “... strict consensus tree resolves only the incorrect relationship (A + B) ...” (Chippindale and Wiens, 1994:284, figure 2 legend). But, the AB node is a constant in all the analyses. In discussing this example, Huelsenbeck *et al.* (1994:290) made the not unreasonable comment that “... the ‘correct’ tree is arbitrarily defined as the most-parsimonious tree for the combined data”. A more reasonable assumption is that, of all the results, the parsimony analysis of the pooled binary data is incorrect because it contains nodes supported by altered data (“optimization” artifacts) and cannot retain the AB node. Proper assessment of matrices seems confounded perpetually by parsimony

analyses of binary data, which are all too often assumed to be the “true” answer. Parsimony might offer too many artifactual nodes for “total evidence” studies to yet be of any real significance. Indeed, it is unknown to what extent artifactual nodes are created in the many “total evidence” cladograms published so far because no one has yet studied the problem. The simple examples above suggest that the problem might be extensive (see also Nelson, 1996; Nelson *et al.* 2003).

## 6. Discussion

Consensus and supertree methodologies, if not all methods of data analysis, can be reduced to determining which data are corroborative, which data are consistent with each other, and which data genuinely conflict. Corroborative and consistent data are relatively straightforward; it is conflict that continues to confound. The problem of conflict, between either components or characters (groups), becomes a little clearer when data are rendered into direct statements of relationships, more specifically as minimal statements of relationship as suites of three-item statements (Nelson and Ladiges, 1991; Nelson and Platnick, 1991; Platnick *et al.*, 1996; Williams, 1996b, 2002). For most conventional consensus and supertree methods, the combination of certain groups is constrained by the binary form of the data and, it appears, little is achieved by tinkering with the methodology. Whereas the “reduced” consensus methods of Adams (1972) and Wilkinson (1994a) deal with portions of the original components (“subcomponents”), neither includes a direct statement of relationship. Both still rely on groups, even if those groups are subgroups of the original component. Without direct statements of relationship, even subgroups have no meaning beyond their membership. Conflict in data is neither Hennig’s dilemma nor Felsenstein’s, but one relevant to binary representation. The problem might best be called the “binary data dilemma” or, perhaps more accurately, the continuing “dilemma of phenetics”. By phenetics, I mean the idea that data might really be “neutral”, when in fact data relevant to consensus and supertrees are associated directly with a particular tree.

It is at this stage that MRP introduces peculiarities of its own. Although it allows the groups within the original components to fragment, it suffers not just from the constraints of binary data, but also from the peculiar effects of optimization, a method that requires all the data to be informative relative to a particular counting method, and thereby having to create artifactual nodes when necessary.

What of “total evidence” and consensus, and their relationship to supertrees? Consider for a moment any data matrix composed of binary

characters drawn from observations on organisms. These data are said to be “shared derived characters”, data indicating (or hypothesizing) particular groups, similar to components. Suppose those data were represented as individual cladograms, such that each matrix was a collection of one-node cladograms instead of being a vast series of 0s and 1s. Because “no one would argue that there is more than one tree of life” (Nelson and Ladiges, 1996:54), each of the one-node cladograms might be “connected” by a single basal, paralogous node (Nelson and Ladiges, 1996:54). By paralogous, I mean that the information in the terminals of each subtree might be duplicated in a manner analogous to individual gene trees, for “without paralogy (in a general sense inclusive of molecular systematics), there is no possibility of either corroboration (Nelson 1994:138) or of conflict” (Nelson and Ladiges, 2001:393). Analysis of these data would then be identical to the general supertree problem, with each subtree (component) specifying the relationships among a set of terminals. Such a view might suggest that the only barrier separating “consensus” from “simultaneous analyses”, and a full appreciation of the supertree problem, is the binary matrix and the algorithms required to “make sense” of the data. Thus, the supertree problem might be seen as a synonym for the general problem relating to systematic data and its relevance.

For the moment, parsimonious analysis of three-item data offers an efficient way of dealing with taxon information because it associates data directly with the implied cladogram. The most efficient summary of all the implied relationships can be found easily and conflict can be dealt with in a more rational way. Three-item data also appear to dispense with the notion that there is a difference between data represented in a matrix or as suites of trees and, by extension, dispenses with the idea that there might be real differences between “character” and “taxonomic” congruence.

In short, it is the issue of binary representation of data, rather than finding the correct or most efficient methodology, that has hindered progress, and, once dispensed with, results from systematic studies might achieve greater precision and accuracy.

## Acknowledgements

I am grateful to Olaf Bininda-Emonds and an anonymous reviewer for pointing out a significant error in an earlier version of this paper.

## References

- ADAMS, E. N. 1972. Consensus techniques and the comparison of taxonomic trees. *Systematic Zoology* 21:390–397.
- BARRETT, M., DONOGHUE, M. J., AND SOBER, E. 1991. Against consensus. *Systematic Zoology* 40:486–493.
- BARRETT, M., DONOGHUE, M. J., AND SOBER, E. 1993. Crusade? A reply to Nelson. *Systematic Biology* 42:216–217.
- BAUM, B. R. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon* 41:3–10.
- BAUM, B. R. AND RAGAN, M. A. 1993. Reply to A. G. Rodrigo's "A comment on Baum's method for combining phylogenetic trees". *Taxon* 42:637–640.
- BAUM, B. R. AND RAGAN, M. A. 2004. The MRP method. In O. R. P. Bininda-Emonds (ed). *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 17–34. Kluwer Academic, Dordrecht, the Netherlands.
- BININDA-EMONDS, O. R. P. AND BRYANT, H. N. 1998. Properties of matrix representation with parsimony analyses. *Systematic Biology* 47:497–508.
- BININDA-EMONDS, O. R. P., GITTLEMAN, J. L., AND STEEL, M. A. 2002. The (super)tree of life: procedures, problems, and prospects. *Annual Review of Ecology and Systematics* 33:265–289.
- CHIPPENDALE, P. T. AND WIENS, J. J. 1994. Weighting, partitioning, and combining characters in phylogenetic analysis. *Systematic Biology* 43:278–287.
- DOYLE, J. J. 1992. Gene trees and species trees: molecular systematics as one-character taxonomy. *Systematic Botany* 17:144–163.
- EERNISSE, D. J. AND KLUGE, A. G. 1993. Taxonomic congruence versus total evidence, and amniote phylogeny inferred from fossils, molecules, and morphology. *Molecular Biology and Evolution* 10:1170–1195.
- ESTABROOK, G. F. 1972. Cladistic methodology: a discussion of the theoretical basis for the induction of evolutionary history. *Annual Review of Ecology and Systematics* 3:427–456.
- FARRIS, J. S. 1973. On comparing the shapes of taxonomic trees. *Systematic Zoology* 22:50–54.
- FARRIS, J. S. AND KLUGE, A. G. 1979. A botanical clique. *Systematic Zoology* 28:400–411.
- FARRIS, J. S. AND KLUGE, A. G. 1997. Parsimony and history. *Systematic Biology* 46:215–218.
- FARRIS, J. S., KLUGE, A. G., AND ECKHART, M. J. 1970. A numerical approach to phylogenetic systematics. *Systematic Zoology* 19:172–189.
- FELSENSTEIN, J. 1982. Numerical methods for inferring evolutionary trees. *Quarterly Review of Biology* 57:379–404.
- FELSENSTEIN, J. 1984. The statistical approach to inferring evolutionary trees and what it tells us about parsimony and compatibility. In T. Duncan and T. F. Stuessy (eds), *Cladistics: Perspectives on the Reconstruction of Evolutionary History*, pp. 169–191. Columbia University Press, New York.
- GOLOBOFF, P. A. AND POL, D. 2002. Semi-strict supertrees. *Cladistics* 18:514–525.
- GORDON, A. D. 1986. Consensus supertrees: the synthesis of rooted trees containing overlapping sets of labeled leaves. *Journal of Classification* 3:335–348.
- HUELSENBECK, J. P., SWOFFORD, D. L., CUNNINGTON, C. W., BULL, J. J., AND WADDELL, P. W. 1994. Is character weighting a panacea for the problem of data heterogeneity in phylogenetic analysis? *Systematic Biology* 43:288–291.

- KITCHING, I. J., FOREY, P. L., HUMPHRIES, C. J., AND WILLIAMS, D. M. 1998. *Cladistics: the Theory and Practice of Parsimony Analysis*. Oxford University Press, Oxford.
- KLUGE, A. G. 1984. The relevance of parsimony to phylogenetic inference. In T. Duncan and T. F. Stussey (eds), *Cladistics: Perspectives on the Reconstruction of Evolutionary History*, pp. 24–38. Columbia University Press, New York.
- KLUGE, A. G. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). *Systematic Zoology* 38:7–25.
- KLUGE, A. G. 1998. Total evidence or taxonomic congruence: cladistics or consensus classification. *Cladistics* 14:151–158.
- KLUGE, A. G. AND WOLF, A. J. 1993. Cladistics: what's in a word? *Cladistics* 9:183–199.
- LORENZEN, S. 1993. The role of parsimony, outgroup analysis, and theory of evolution in phylogenetic systematics. *Zeitschrift für Zoologische Systematik und Evolutionsforschung* 31:1–20.
- MIYAMOTO, M. M. 1985. Consensus cladograms and general classifications. *Cladistics* 1:186–189.
- MIYAMOTO, M. M. AND FITCH, W. M. 1995. Testing species phylogenies and phylogenetic methods with congruence. *Systematic Biology* 44:64–76.
- MOORE, J. M. 2001. [Review of "Disseminating Darwin"]. *Books and Culture* 7:36.
- MORSE, J. C. AND WHITE, D. F., JR. 1979. A technique for analysis of historical biogeography and other characters in comparative biology. *Systematic Zoology* 28:356–365.
- NELSON, G. J. 1979. Cladistic analysis and synthesis: principles and definitions, with a historical note on Adanson's *Families des Plantes* (1763–1764). *Systematic Zoology* 28:1–21.
- NELSON, G. J. 1989. Cladistics and evolutionary models. *Cladistics* 5:275–289.
- NELSON, G. J. 1993. Why crusade against consensus? A reply to Barrett, Donoghue, and Sober. *Systematic Biology* 42:215–216.
- NELSON, G. J. 1994. Homology and systematics. In B. K. Hall (ed.), *Homology: the Hierarchical Basis of Comparative Biology*, pp. 101–149. Academic Press, San Diego.
- NELSON, G. J. 1996. Nullius in Verba. *Journal of Comparative Biology* 1:141–152.
- NELSON, G. J. In press. Cladistics: its arrested development. In D. M. Williams and P. L. Forey (eds), *Milestones in Systematics*. Taylor & Francis, London.
- NELSON, G. J. AND LADIGES, P. Y. 1991. Three-area statements: standard assumptions for biogeographic analysis. *Systematic Zoology* 40:470–485.
- NELSON, G. J. AND LADIGES, P. Y. 1994. Three-item consensus: empirical test of fractional weighting. In R. W. Scotland, D. J. Siebert, and D. M. Williams (eds), *Models in Phylogeny Reconstruction*, pp. 193–209. Clarendon Press, Oxford.
- NELSON, G. J. AND LADIGES, P. Y. 1996. Paralogy in cladistic biogeography and analysis of paralogy-free subtrees. *American Museum Novitates* 3167:1–58.
- NELSON, G. J. AND LADIGES, P. Y. 2001. Gondwana, vicariance biogeography and the New York School revisited. *Australian Journal of Botany* 49:389–409.
- NELSON, G. J. AND PLATNICK, N. I. 1980. Multiple branching in cladograms: two interpretations. *Systematic Zoology* 29:86–91.
- NELSON, G. J. AND PLATNICK, N. I. 1981. *Systematics and Biogeography: Cladistics and Vicariance*. Columbia University Press, New York.
- NELSON, G. J. AND PLATNICK, N. I. 1991. Three-taxon statements: a more precise use of parsimony? *Cladistics* 7:351–366.
- NELSON, G. J., WILLIAMS, D. M., AND EBACH, M. C. 2003. A question of conflict: three item and standard parsimony compared. *Systematics and Biodiversity* 2:145–149.

- NIXON, K. C. AND CARPENTER, J. M. 1996. On consensus, collapsibility, and clade concordance. *Cladistics* 12:305–321.
- PAGE, R. D. M. 1987. Graphs and generalized tracks: quantifying Croizat's panbiogeography. *Systematic Zoology* 36:1–12.
- PAGE, R. D. M. 1989. Comments on component-compatibility in historical biogeography. *Cladistics* 5:167–182.
- PAGE, R. D. M. 1990a. Component analysis: a valiant failure? *Cladistics* 6:119–136.
- PAGE, R. D. M. 1990b. Tracks and trees in the Antipodes: a reply to Humphries and Seberg. *Systematic Zoology* 39:288–299.
- PAGE, R. D. M. 1994. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Systematic Biology* 43:58–77.
- PATTERSON, C. 1982. Morphological characters and homology. In K. A. Joysey and A. E. Friday (eds), *Problems of Phylogenetic Reconstruction*, pp. 21–74. Academic Press, London.
- PATTERSON, C. 1988. Homology in classical and molecular biology. *Molecular Phylogenetics and Evolution* 5:603–625.
- PISANI, D. AND WILKINSON, M. 2002. Matrix representation with parsimony, taxonomic congruence, and total evidence. *Systematic Biology* 51:151–155.
- PLATNICK, N. I. 1993. Character optimization and weighting: differences between the standard and three-taxon approaches to phylogenetic inference. *Cladistics* 9:267–272.
- PLATNICK N. I., HUMPHRIES, C. J., NELSON, G. J., AND WILLIAMS, D. M. 1996. Is Farris optimization perfect? *Cladistics* 12:243–252.
- RAGAN, M. A. 1992a. Matrix representation in reconstructing phylogenetic relationships among the eukaryotes. *BioSystems* 28:47–55.
- RAGAN, M. A. 1992b. Phylogenetic inference based on matrix representation of trees. *Molecular Phylogenetics and Evolution* 1:53–58.
- RODRIGO, A. G. 1993. A comment on Baum's method for combining phylogenetic trees. *Taxon* 42:631–636.
- RODRIGO, A. G. 1996. On combining cladograms. *Taxon* 45:267–274.
- ROSS, H. A. AND RODRIGO, A. G. 2004. An assessment of matrix representation with compatibility in supertree construction. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 35–63. Kluwer Academic, Dordrecht, the Netherlands.
- SANDERSON, M. J., PURVIS, A., AND HENZE, C. 1998. Phylogenetic supertrees: assembling the trees of life. *Trends in Ecology and Evolution* 13:105–109.
- SCOTLAND, R. W. 1992. Cladistic theory. In P. L. Forey, C. J. Humphries, I. J. Kitching, R. W. Scotland, D. J. Siebert, and D. M. Williams. *Cladistics: A Practical Course in Systematics*, pp. 3–13. Oxford University Press, Oxford.
- SCOTLAND, R. W. 1997. Parsimony neither maximizes congruence nor minimizes incongruence or homoplasy. *Taxon* 46:743–746.
- SCOTLAND, R. W. 2000. Homology, coding and three-taxon statement analysis. In R. W. Scotland and T. Pennington (eds), *Homology and Systematics*, pp. 145–182. Taylor and Francis, London.
- SNEATH, P. H. A. 1988. The phenetic and cladistic approaches. In D. L. Hawksworth (ed.), *Prospects in systematics*, pp. 252–273. Clarendon Press, Oxford.
- SNEATH, P. H. A. 1995. Thirty years of numerical taxonomy. *Systematic Biology* 44:281–298.
- SOKAL, R. R. AND SNEATH, P. H. A. 1963. *Principles of Numerical Taxonomy*. W. H. Freeman, San Francisco.

- WÄGELE, J. W. 1994. Review of methodological problems of “computer cladistics” exemplified with a case study on isopod phylogeny (Crustacea: Isopoda). *Zeitschrift für zoologische Systematik und Evolutionsforschung* 32:81–107.
- WILKINSON, M. 1994a. Common cladistic information and its consensus representation: reduced Adams and reduced cladistic consensus trees and profiles. *Systematic Biology* 43:343–368.
- WILKINSON, M. 1994b. Three-taxon statements: when is a parsimony analysis also a clique analysis? *Cladistics* 10:221–223.
- WILKINSON, M. 1994c. The permutation method and character compatibility. *Systematic Biology* 43:274–277.
- WILKINSON, M AND THORLEY, J. L. 1998. Reduced supertrees. *Trends in Ecology and Evolution* 13:283.
- WILKINSON, M, THORLEY, J. L., LITTLEWOOD, D. T. J., AND BRAY, R. A. 2001. Towards a phylogenetic supertree of Platyhelminthes? In D. T. J. Littlewood and R. A. Bray (eds), *Interrelationships of the Platyhelminthes*, pp. 292–301. Taylor and Francis, London.
- WILLIAMS, D. M. 1994. Combining trees and combining data. *Taxon* 43:449–453.
- WILLIAMS, D. M. 1996a. Characters and cladograms. *Taxon* 45:275–283.
- WILLIAMS, D. M. 1996b. Fossil species of the diatom genus *Tetraclcyclus* (Bacillariophyta, ‘ellipticus’ group): morphology, interrelationships and the relevance of morphogenesis to phylogeny. *Philosophical Transactions of the Royal Society of London Series B* 351:1759–1782.
- WILLIAMS, D. M. 2002. Parsimony and precision. *Taxon* 51:143–149.
- WILLIAMS, D. M. In press. Homology and homologues, cladistics and phenetics: 150 years of progress. In D. M. Williams and P. L. Forey (eds), *Milestones in Systematics*. Taylor & Francis, London.
- WILLIAMS, D. M. AND HUMPHRIES, C. J. 2003a. Component coding, three-item coding and consensus methods. *Systematic Biology* 52:255–259.
- WILLIAMS, D. M. AND HUMPHRIES, C. J. 2003b. Homology and the evolution of characters. In T. Stuessy, E. Hörandl, and V. Mayer (eds), *Deep Morphology: Toward a Renaissance of Morphology in Plant Systematics*, pp. 119–130. Königstein, Koeltz.
- WILLIAMS, D. M. AND SIEBERT, D. J. 2000. Characters, homology and three-item analysis. In R. W. Scotland and T. Pennington (ed.), *Homology and Systematics*, pp. 183–208. Taylor and Francis, London.

## Chapter 19

# A MOLECULAR SUPERTREE OF THE ARTIODACTYLA

Annette S. Mahon

**Abstract:** Despite the size of the order and the conservation importance of many of its members, no complete species-level phylogeny of extant artiodactyls (*sensu stricto*) exists. Matrix Representation with Parsimony, which has been used already in reconstructions of primate and carnivore phylogeny, was used to build a supertree of the order. Owing to a lack of data, only 171 of the 220 extant species could be included in the analysis. Forty-eight molecular source trees contributed to building a supertree, with a current (morphological) taxonomy used to provide a backbone. The resulting supertree largely reflects a consensus of recent molecular work; however, resolution of the tree varies across families reflecting areas of current uncertainty. A discussion of the structure of the tree, and of its possible limitations, is presented.

**Keywords:** Artiodactyla; MRP; parsimony ratchet; supermatrix; supertree

### 1. Introduction

Only fragmentary information is available concerning the evolutionary relationships of the order Artiodactyla, with much of it focusing on a few major families and tribes. Composite trees have been used in studies of co-adaptation (Brashares *et al.*, 2000) and reproductive strategies (Sæther and Gordon, 1994), but these have been constructed informally. The traditional view that Artiodactyla constitutes a monophyletic group (Eisenberg, 1981; Prothero, 1993) is based primarily on morphological cranial characters. Cetacea (whales) were thought to be related to the extinct carnivorous group Mesonychia based on dentary characters (Gingerich *et al.*, 1983). However,

in the past ten years, a large amount of molecular evidence has suggested that cetaceans are closely related to artiodactyls. Some studies take this further to suggest that Cetacea form a clade within Artiodactyla to form the sister group to Hippopotamidae (e.g., Graur and Higgins, 1994; Irwin and Arnason, 1994; Smith *et al.*, 1996; Shimamura *et al.*, 1997), making artiodactyls paraphyletic. Other evidence in support of this relationship came from milk casein genes and  $\gamma$ -fibrinogen sequences (Gatesy *et al.*, 1996; Gatesy, 1998). Although the position of Cetacea within the artiodactyls is accepted increasingly, the wider project of which this supertree forms part does not include cetaceans, and they have been omitted from this analysis.

This analysis aims to provide a summary of available molecular information about Artiodactyla, which will be of systematic interest and be relevant to the study of artiodactyls (e.g., life-history analysis or for conservation purposes).

## 2. Method

I used Matrix Representation with Parsimony (MRP; Baum, 1992; Ragan, 1992) to combine the trees. MRP is described more fully elsewhere (e.g., Baum and Ragan, 2004), but can be summarized briefly as follows: for each clade in a source tree, taxa that are members of that clade are coded as 1, taxa absent from that clade are coded as 0, and taxa missing from individual source trees are coded as “missing data” (using ?). Trees are rooted using a hypothetical all-zero outgroup. By reducing each tree to a binary matrix, a homoplasy-free representation of its structure can be obtained. The trees used in MRP can be derived from a wide range of (incompatible) data types, often meaning that more data can be included in the analysis. Source trees with different species can be also combined; all that is required is that a tree has at least one species in common with at least two other source trees (Sanderson *et al.*, 1998). For a review of the difficulties of, and usefulness of, combining trees, see Bininda-Emonds and Bryant (1998).

There is an ongoing debate about the relative value of supermatrix and supertree methods (Gatesy *et al.*, 2002; Bininda-Emonds *et al.*, 2003). In this study, the potential to use a supermatrix approach existed because only molecular studies were analyzed. However, many different genes and proteins were examined, and these are difficult to examine simultaneously using the appropriate models of evolution for each. In particular, adequate maximum likelihood models for allozymes and proteins are debatable, which would limit the analysis to a parsimony approach. The advantage of MRP lies in its ability to combine data sources where each data set was analyzed separately under an appropriate evolutionary model in the source studies

(Bininda-Emonds *et al.*, 2002, 2003). Furthermore, MRP shows good performance in simulation, providing a good approximation of both the total evidence solution and of the model tree in most cases (Bininda-Emonds and Sanderson 2001).

MRP, despite some problems (e.g., the creation of novel clades and effects of very incongruent source trees; Bininda-Emonds and Bryant, 1998; Bininda-Emonds and Sanderson, 2001), has now been used to obtain species-level supertrees for the primates (Purvis, 1995; Purvis and Webster, 1999), carnivores (Bininda-Emonds *et al.*, 1999), bats (Jones *et al.*, 2002), a genus-level tree of the Dinosauria (Pisani *et al.*, 2001), and an ordinal / familial supertree of all mammals (Liu *et al.*, 2001), among others (see also Baum and Ragan, 2004).

## 2.1 Independence of source trees

Like other phylogenetic-reconstruction techniques, MRP assumes that the source data are independent. However, as Gatesy *et al.* (2002) show, non-independence among potential source studies can be pervasive. In particular, where molecular sequences are used repeatedly in successive analyses, the resulting trees cannot be said to be independent because part of their source data has been reused.

This study has not used an explicit system as given elsewhere (Bininda-Emonds *et al.*, 2004); however, attempts have been made to maintain the independence of source trees. For source trees that are also supertrees, the original literature was referred to if possible. Also, as a precaution, all source trees were cross-referenced against each other in an attempt to ensure that no information is included more than once and therefore given undue weight. This approach had limited success. With molecular studies, most work has been done on particular sequences (usually 12S / 16S rDNA or cytochrome *b*) that have been used repeatedly (e.g., Chikuni *et al.*, 1995; Irwin *et al.*, 1991). Therefore, although every effort was made to avoid repeating information, some duplication is inevitable. However, although sequences might have been re-used, differences in alignment or addition of extra species might mean that the methods or the results are not the same as previous analyses (see Bininda-Emonds *et al.*, 2002, 2003, 2004). This is true especially when old and new data are combined (e.g., Gatesy *et al.*, 1992, 1997; Hassanin and Douzery, 1999a, b).

## 2.2 Taxonomy

The importance of maintaining a single reference taxonomy has already been noted (Bininda-Emonds *et al.*, 2004), and the taxonomy used in this study

follows that of *Mammal Species of the World* (Grubb, 1993). Where defunct species names were used in source trees, these have been changed to reflect those of Grubb (1993). Exceptions include species that have been discovered since 1993, such as *Pseudoryx nghetinhensis* (Hassanin and Douzery, 1999a) and four species of muntjac deer, *Muntiacus putaoensis*, *M. rooseveltorum*, *M. truongsonensis*, and *M. vuquangensis* (Amato *et al.*, 1999, 2000; Schaller and Vrba, 1996). By contrast, some recently extinct species such as *Bubalus mephistopheles* that are listed by Grubb (1993) have not been used, although *Phacochoerus aethiopicus* is included following Theimer and Keim (1998).

As discussed above, Cetacea are often included as a monophyletic group within Artiodactyla. For operational reasons, they are not included here. Thus, when dealing with source trees that included cetaceans, I have excised them and coded the reduced tree in the matrix.

Subspecies were also not used in the reconstruction. The importance of subspecies in mammalian taxonomy is the subject of a continuing debate in conservation biology, particularly as decisions are made about safeguarding habitats or individual populations (e.g., Molina and Molinari, 1998; Vázquez and Gittleman, 1998; Su *et al.*, 1999). However, the taxonomic status of some subspecies is still debated (e.g., numerous species within Cervidae; Lowe and Gardiner, 1989; Cook *et al.*, 1999). As a result, where subspecies or races are included in the source data, the node has been collapsed to species level. This procedure allows the possibility that the entities created might not be monophyletic (see Bininda-Emonds *et al.*, 2004).

### 2.3 Data collection

Potential source trees were identified initially by searching the Web of Science using a variety of search terms. Initial searches, and searches within these results, produced a primary list of approximately 150 source papers. These were scanned, sometimes in abstract form, for usefulness, and a total of 48 molecular source studies were used to build a supertree. Source data included cytochrome *b* and cytochrome *c*, 12S and 16S rDNA, and allozyme studies. Many of the source trees are small, with 18 of the source trees containing less than 18 species. Many of the smaller trees concentrate on specific families, giving a greater resolution than might otherwise have been obtained. However, this can also be a disadvantage because smaller trees often supply the *only* detailed information for particular families (e.g., Stanley (1994) for the camelids). A current taxonomy (Grubb, 1993) was used to provide a backbone to seed the analysis by coding the tree implied by the taxonomy (see Bininda-Emonds and Sanderson, 2001).

Initially, the goal was to include all 220 extant species, but no molecular source data were available for some species. I attempted initially to place

such taxa at the lowest possible taxonomic level, usually that of family or subfamily, to include them in the supertree without making any additional assumptions about their relationships. These species were coded into the supertree using taxonomic information from Grubb (1993) only; all other characters for these species were coded as missing data. This should have resulted in a situation where each family / subfamily consisted of a set of resolved species and a bush of species for which there was no information except that they belong to that family. It is well known that MRP reacts poorly to characters that do not overlap on several sources because they can be inserted equally parsimoniously almost anywhere on the tree. Herein, although the species remained within their respective families, they often either inserted apparently randomly or as a ladderized (not bush-like) group. Because both of these outcomes give a false impression of tree structure, such cases were removed before all other analyses were undertaken. Therefore, 49 species for which no molecular data was available were not included in all subsequent analyses. The families Suidae and Cervidae, and the subfamily Antilopinae have a disproportionately large number of missing taxa for the number of species in each.

The supertree contains 171 taxa (77%) of extant artiodactyl species. Source trees were coded into a matrix of 702 characters and analyzed using PAUP\* 4.0b10 (Swofford, 2002). TreeView v1.5 (Page, 1996) was used to display the resulting trees. Table 1 summarizes the source trees and groups included.

It is worth noting that a relatively small number of source trees (48) were used to construct this phylogeny. This is in contrast with 112 for Primates (Purvis, 1995), 177 for Carnivora (Bininda-Emonds *et al.*, 1999), and 430 for placental mammals (Liu *et al.*, 2001). There are several reasons for this. The artiodactyls have been studied less intensively than either the primates or carnivores. Supertree studies of these latter groups also selected a wider range of source trees for inclusion. For example, trees derived from non-cladistic analysis or phenograms were used by Purvis, and Bininda-Emonds *et al.* used all available data. Springer and de Jong (2001) noted that Liu *et al.* (2001) included source trees that cannot be reconstructed from the original matrices, and Gatesy *et al.* (2002; also Gatesy and Springer, 2004) were also critical of data collection for MRP in general. To avoid some of these problems, all source trees used in the artiodactyl supertree were molecular and post-1988, and it is hoped that their smaller number is balanced by their increased quality. Where a consensus tree was available in the source papers, it was used in the supertree. If a consensus tree was not given (or was unusable because it contained morphological data), component trees were used instead.

*Table 1.* Summary of source data used in supertree construction. Numbers in the left hand columns refer to the start position of the data source in the matrix. Figure and page numbers refer to the original reference.

|     | Study                           | Taxon          | Data Source            | Fig. | Page  |
|-----|---------------------------------|----------------|------------------------|------|-------|
| 1   | Grubb (1993)                    | All            |                        |      |       |
| 28  | Allard <i>et al.</i> (1992)     | Bovidae        | 12S and 16S rDNA       | 2B   | 3974  |
| 42  | Cronin <i>et al.</i> (1996)     | Pecora         | $\kappa$ -casein       | 3    | 306   |
| 65  | Essop <i>et al.</i> (1997)      | Bovidae        | mtDNA                  | 2    | 382   |
| 76  | Groves and Shields (1996)       | Caprinae       | cyt b                  | 3    | 472   |
| 86  | Georgiadis <i>et al.</i> (1990) | Bovidae        | allozyme               | 4    | 2142  |
| 106 | Hassanin and Douzery (1999a)    | Bovidae        | cyt b                  | 2A   | 897   |
| 132 | Hassanin and Douzery (1999b)    | Bovidae        | cyt b                  | 4    | 238   |
| 170 | Janecek <i>et al.</i> (1996)    | Bovinae        | cyt c                  | 5A   | 115   |
| 181 | Matthee and Robinson (1999)     | Bovidae        | cyt b                  | 5    | 44    |
| 219 | Miyamoto <i>et al.</i> (1989)   | Bovini         | mtDNA                  | 1A   | 345   |
| 224 | Pitra <i>et al.</i> (1997)      | Bovini         | nuclear DNA            | 2B   | 595   |
| 234 | Randi <i>et al.</i> (1998)      | Cervidae       | cyt b                  | 2    | 798   |
| 252 | Schreiber <i>et al.</i> (1999)  | Bovidae        | cyt b                  | 1    | 171   |
| 261 | Su <i>et al.</i> (1999)         | <i>Moschus</i> | cyt b                  | 2A   | 247   |
| 288 | Stanley <i>et al.</i> (1994)    | Camelidae      | cyt b                  | 1    | 3     |
| 294 | Theimer and Keim (1998)         | Peccaries      | cyt b                  | 1    | 568   |
| 298 | Vassart <i>et al.</i> (1995)    | <i>Gazella</i> | chromosome             | 12B  | 225   |
| 308 | Van Vuuren and Robinson(2001)   | Cephalopinae   | cyt b                  | 1A   | 416   |
| 329 | Wang and Lan (2000)             | Muntiacinae    | chromosome             | 11   | 328   |
| 337 | Hassanin <i>et al.</i> (1998)   | Caprinae       | cyt b                  | 2A   | 226   |
| 367 | Polzhein and Strobeck (1998)    | Cervinae       | mtDNA                  | 2    | 254   |
| 376 | Hartl <i>et al.</i> (1990)      | Caprini        | protein                | 4    | 180   |
| 383 | Wall <i>et al.</i> (1992)       | Bovinae        | rDNA                   | 4B   | 273   |
| 394 | Birungi and Arctander (2001)    | Reduncini      | cyt b                  | 3    | 136   |
| 406 | Matthee <i>et al.</i> (2001)    | Artiodactyla   | nuclear DNA            | 2    | 377   |
| 429 | Emerson and Tate (1993)         | Cervinae       | protein                | 3    | 270   |
| 434 | Randi <i>et al.</i> (1991)      | Caprinae       | allozyme               | 2D   | 284   |
| 437 | Miyamoto <i>et al.</i> (1990)   | Antlered Deer  | mtDNA                  | 2    | 6130  |
| 440 | Van Vuuren <i>et al.</i> (2001) | Cephalopinae   | 12S rDNA               | 2A   | 418   |
| 454 | Hassanin and Douzery (2003)     | Ruminantia     | nuclear and mt markers | 5    | 233   |
| 479 | Randi <i>et al.</i> (1996)      | Suiformes      | cyt b                  | 3    | 179   |
| 483 | Gatesy (1998)                   | All            | $\gamma$ -fibrinogen   | 9    | 75    |
| 493 | Gatesy and Arctander (2000)     | Bovidae        | $\alpha$ -lactalbumin  | 5    | 527   |
| 505 | Nikaido <i>et al.</i> (1999)    | All            | SINE / LINEs           | 7    | 10265 |

Table 1. Continued

|     | Study                           | Taxon        | Data Source          | Fig. | Page |
|-----|---------------------------------|--------------|----------------------|------|------|
| 509 | Gatesy (1998)                   | Bovidae      | protamine P1         | 14   | 80   |
| 519 | Madsen et al. (2001)            | All          | nuclear and mt genes | 1B   | 611  |
| 522 | Kleinedam et al. (1999)         | All          | ribonuclease         | 2    | 363  |
| 525 | Gatesy (1998)                   | All          | $\kappa$ -casein     | 12   | 78   |
| 557 | Gatesy (1998)                   | All          | cyt b                | 10   | 76   |
| 589 | Murphy et al. (2001)            | All          | nuclear and mtDNA    | 1    | 614  |
| 594 | Gatesy and Arctander (2000)     | Bovidae      | 12S and 16S rDNA     | 2    | 525  |
| 632 | Gatesy and Arctander (2000)     | Bovidae      | $\beta$ -casein      | 4    | 527  |
| 649 | Gatesy (1998)                   | All          | $\beta$ -casein      | 13   | 79   |
| 677 | Douzery and Randi (1997)        | Cervidae     | mt control region    | 6    | 1161 |
| 687 | Grobler and van der Bank (1995) | Alcelaphinae | allozyme             | 2    | 306  |
| 689 | Beintema et al. (2003)          | All          | protein              | 4    | 24   |
| 695 | Montlegard et al. (1998)        | All          | cyt b, 12S rDNA      | 1A   | 530  |

Three levels of source data were included in the supertree: 1) within families (which generally had a relatively small number of species), 2) across families (numerous species), and 3) higher-level trees (which might have only one species for each family / subfamily). The smaller within-family trees create islands of information, but the remaining levels of source data bridge this. An analysis of the relative impact of each type of source tree on the structure of the final supertree would be very valuable.

Fossil species have, of course, been excluded from the supertree, but an integration of morphological (both extinct and extant) data could be very fruitful (for molecular and morphology comparisons at ordinal level, see Shoshani and McKenna, 1998). Although it is increasingly common to include all data in an analysis, it was felt that a separate analysis of the molecular data would be beneficial for comparison with previously published morphological work.

## 2.4 The parsimony ratchet

The size of the order Artiodactyla, and of some of the constituent families, creates difficulties, even when one uses heuristic search methods. For 171 species, as herein, the tree space to be searched is large and an optimal result might not be found despite large amounts of computer time and memory. One solution to this problem is a compartmentalization approach, which deals with smaller subsets of data (e.g., Bininda-Emonds *et al.*, 1999). This is not ideal because it forces assumptions to be made concerning monophyly

of families and genera, and concerning how the families relate to one other. There is also the potential problem that source trees can conflict with the assumptions of monophyly (and would need to be discarded) because of, among other causes, problematic taxa that jump between families in different source trees depending on the type of data used.

The parsimony ratchet (Nixon, 1999) is a search strategy that allows large matrices to be analyzed in a reasonable period of time, and has been applied in a previous large-scale supertree analysis (Jones *et al.*, 2002). This method is much faster than a normal heuristic search, and has the advantage that it is less likely to be trapped on suboptimal islands because the search jumps from one section of tree space to another. All 171 species could therefore be analyzed simultaneously, although each family and subfamily within Bovidae is treated separately below for ease of discussion. Searches were performed using PAUP\* v4b10 (Swofford, 2002), with the ratchet instructions created using the Perl script perlRat.pl (available from <http://www.tierzucht.tum.de/Bininda-Emonds>). The format of the ratchet was to run 10 batches of 500 replicates each, and to save all trees to a single file, following Nixon's (1999) suggestion that multiple shorter searches reduce the possibility of the ratchet becoming trapped on suboptimal islands. A random sample of 25% of characters was reweighted at each iteration. All saved trees were then read into memory and used as starting trees for a heuristic search (TBR branch swapping, multiple trees retained) after suboptimal trees were discarded. A maximum of 10 000 trees was retained. A similar approach was used by Jones *et al.* (2002). Because the ratchet is a more efficient method than a normal heuristic search, 10 000 trees are likely to be an effective representation of the tree space (Nixon, 1999).

Bremer supports (Bremer, 1988) have been reported for MRP supertrees traditionally (e.g., Bininda-Emonds *et al.*, 1999; Liu *et al.*, 2001; Jones *et al.*, 2002), and although their applicability to this type of analysis has been queried (Pisani *et al.*, 2002), they are given here as an index of support. Bremer supports were determined using the multiple parsimony ratchet searches, each of which was constrained to find trees that did not contain a specified node. The format of each ratchet was 20 batches of 100 iterations each with no subsequent search.

### 3. Results and discussion

The ratchet search described above yielded 10 000 equally most parsimonious trees (MPTs), each of 860 steps. Owing to space limitations, the overall supertree is shown at the family level only (Figure 1) with species-level relationships within the families / subfamilies presented in

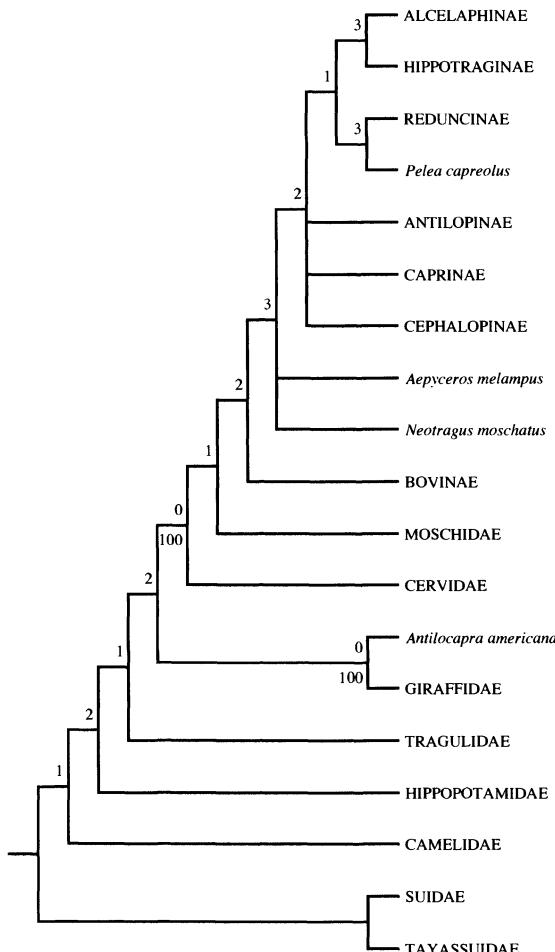


Figure 1. Topology of complete supertree (90% majority rule consensus tree). Antilopinae form a monophyletic group with 96% support. All other family / subfamily nodes are supported at 100%.

Figures 2–8. A single supertree with all species is available on request from the author. The different trees were usually summarized using strict consensus. Both Antilopinae and Cervidae were completely unresolved with strict consensus, so 90% majority-rule supertrees for these groups are shown. Unless otherwise indicated, the strict-consensus and majority-rule supertrees are identical. Numbers above the branches are Bremer support values, which are low generally, and, where applicable, numbers below branches are 90% majority-rule percentages. All branch lengths are arbitrary.

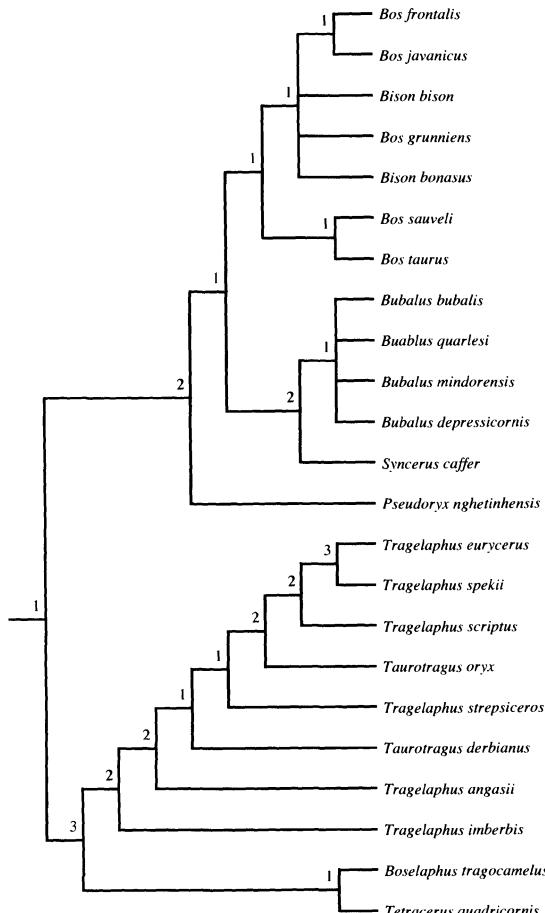


Figure 2. Supertree of Bovinae. Species not included: *Tragelaphus buxtoni*.

### 3.1 Bovidae and Bovinae (Figure 2)

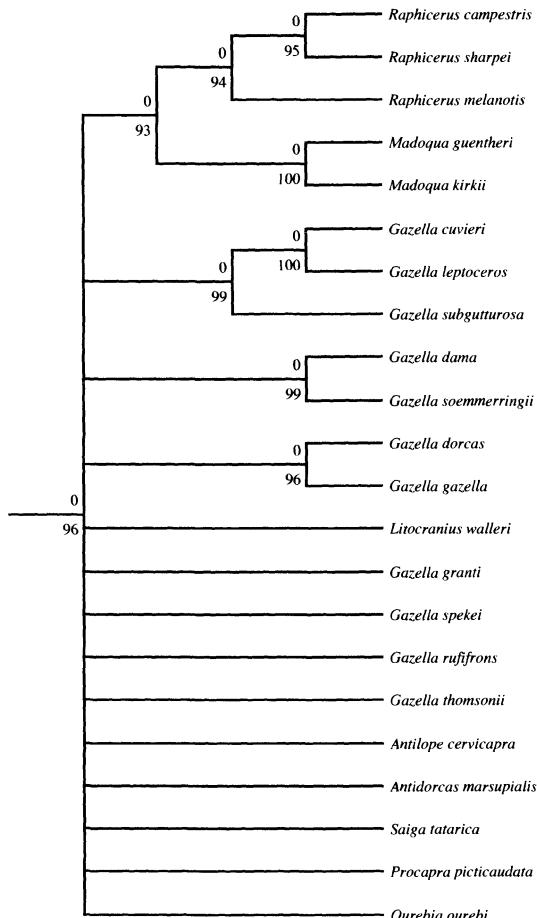
Despite some suggestions to the contrary (Gatesy *et al.*, 1992; Essop *et al.*, 1997), Bovidae is regarded generally as a monophyletic group (Allard *et al.*, 1992). However, there is general disagreement over the branching pattern within the family (e.g., Georgiadis *et al.*, 1990; Allard *et al.*, 1992; Gatesy *et al.*, 1992). This disagreement is exacerbated by the use of both tribes and families as basic divisions in the literature. In the supertree, there is a clear division between Bovinae (tribes Tragelaphini, Boselaphini, and Bovini) and the remainder of Bovidae. This division is well supported within the source data and independently (Robinson *et al.*, 1998; but see Georgiadis, 1990).

The subfamily Bovinae (Figure 2) is one of the best-resolved subfamilies within artiodactyls, and there are many more source data available about the relationships of Bovinae than for all the other groups. In the supertree, Bovinae are a very well-resolved group, reflecting the general consensus among the source trees.

### 3.2 Antilopinae (Figure 3)

In the supertree, Antilopinae is completely unresolved with strict consensus. This is a large subfamily within artiodactyls of considerable conservation importance: several of the 36 species are listed in the 2002 IUCN Red List of Threatened Species (<http://www.redlist.org>), ranging from conservation-dependent species to the critically endangered *Procapra przewalskii*. Therefore, it is surprising how fragmentary the phylogenetic information about Antilopinae is. Most of the available information concentrates on the gazelles (e.g., Vassart *et al.*, 1995), and the problem is exacerbated by the lack of a comprehensive molecular phylogeny that includes all the main genera (*Gazella*, *Madoqua*, *Neotragus*, *Procapra*, and *Raphicerus*), and shows how they relate to one other.

When MPTs are resolved using 90% majority rule (Figure 3), some structure does appear: *Gazella* is not clearly monophyletic, and *Raphicerus* and *Madoqua* are sister groups. Groves (2000) conducted a morphological analysis of the group and concluded that *Gazella* is polyphyletic with respect to *Antilope*. Further suggestions that *Ammendorcas* and *Antidorcas* form a clade and that *Procapra* and *Prodorcas* group together cannot be commented on because one or more of these species was missing from the analysis. *Saiga tatarica*, which has been placed with the caprids in the past, groups here within the antelopes in agreement with the morphological analysis of Gentry (1978a, 1990). The more resolved majority-rule tree suggests that differences within the available data are causing the collapse under strict consensus. Although there is some genuine conflict between the source trees, different taxon combinations are often used, particularly in the larger interfamily tree, which leads to poor resolution within the strict consensus tree. Restricted taxonomic overlap has been shown to affect supertree construction negatively (Bininda-Emonds and Sanderson, 2001). Morphological data, which is not considered here, might also help to resolve the group. Relationships within antelopes should be regarded with a certain amount of caution when considering the supertree as a whole, although the situation is not as bad as a first glance at the consensus tree would suggest. However, this is a group that would benefit still from a comprehensive molecular investigation.



*Figure 3.* Supertree of Antilopinae (90% majority rule consensus). Species not included: *Ammodorcas clarke*, *Gazella arabicai*, *Gazella bennettii*, *Gazella bilkis*, *Gazella rufina*, *Gazella saudiya*, *Dorcatragus megalotis*, *Madoqua piacentinii*, *Madoqua saltiana*, *Neotragus batesi*, *Neotragus pygmeus*, *Procapra gutturosa*, and *Procapra przewalskii*.

### 3.3 Cephalophinae (Figure 4)

The duikers are a speciose group of generally forest-dwelling antelope. The exception is *Sylvicapra grimmia*, which is larger, with more upright horns, and lives more open areas (Gentry, 1990). Because of the difference in size and habitat, *Sylvicapra grimmia* is usually held to be the sister group to the remainder of Cephalophinae. Here, it is included within *Cephalophus*,

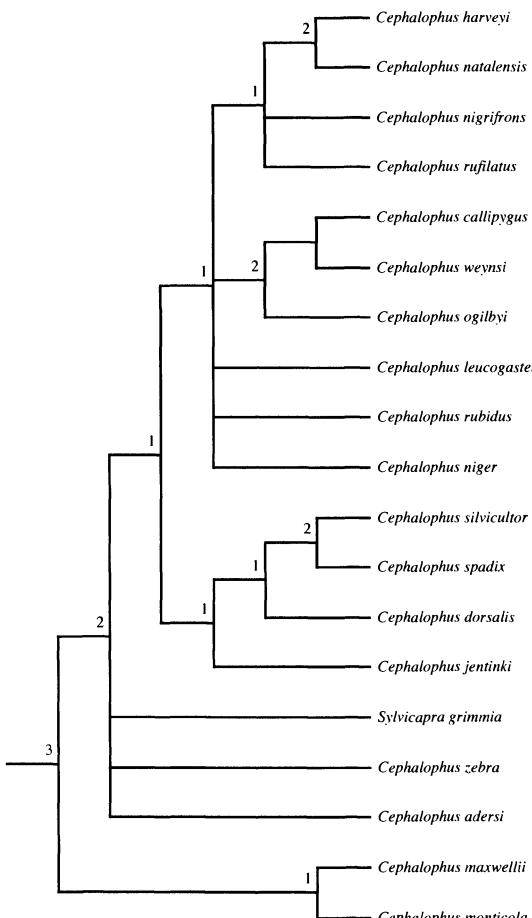


Figure 4. Supertree of Cephalophinae.

probably because of source data that place it close to either *C. maxwellii* or *C. monticola*. Most molecular work on duikers has been carried out by van Vuuren and Robinson (2001) and certain groupings seem robust: *Cephalophus callipygus*, *C. weynsi*, and *C. ogilbyi*; *C. harveyi*, *C. natalensis*, *C. nigrifrons*, and *C. rufilatus*; and *C. silvicultor*, *C. spadix*, and *C. dorsalis*. On the basis of karyotype comparisons, Robinson *et al.* (1996) suggested that *C. maxwellii* and *C. monticola* are sister taxa. The groupings above appear in the supertree, but relationships of the remaining species are less well understood.

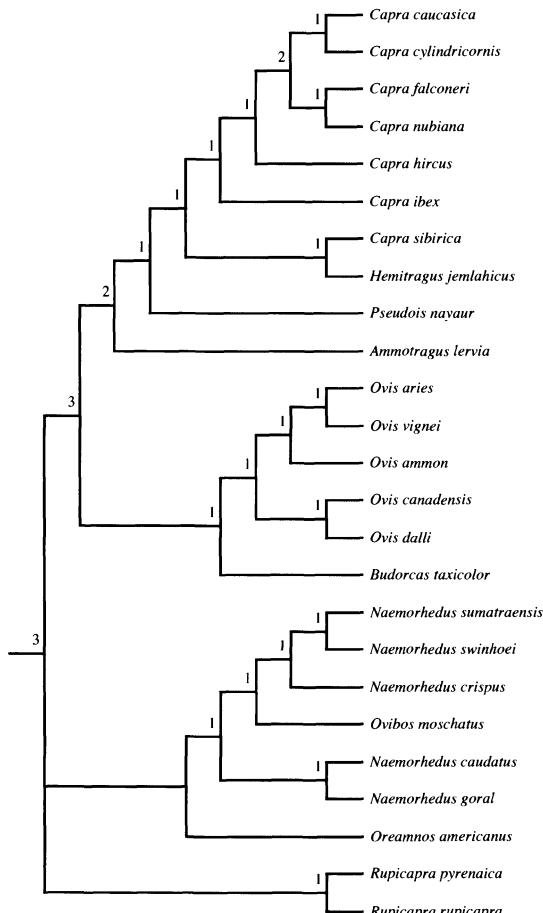
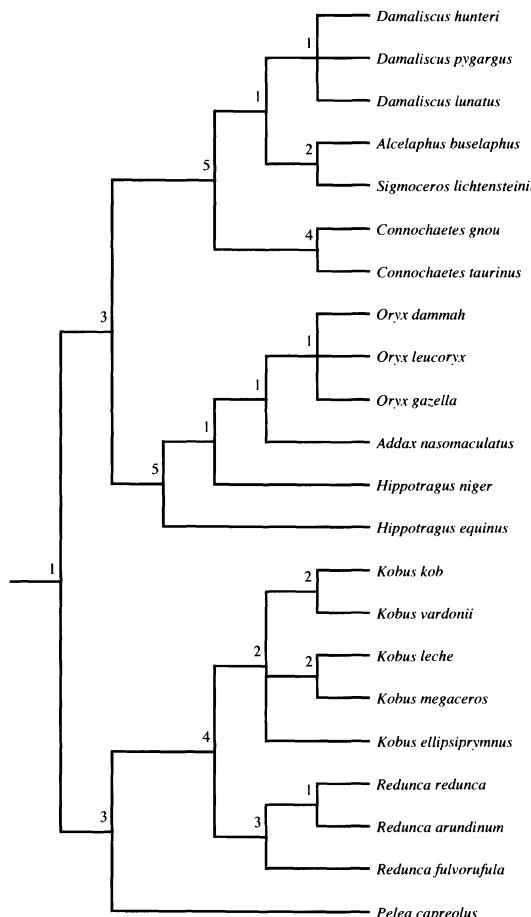


Figure 5. Supertree of Caprinae.

### 3.4 Caprinae (Figure 5)

Two species are particularly problematic when looking at Caprinae: the saiga, *Saiga tatarica*, and chiru, *Pantholops hodgsonii*. Both have caused controversy in the past and have been grouped with caprids (within the tribe Saigini; e.g., Hassanin *et al.*, 1998) or within Antilopinae (e.g., Gatesy *et al.*, 1997). In the supertree, *Pantholops* was the sister group to the caprids, in contrast to morphological evidence (Gentry, 1978a). However, the Bremer support for this node was zero, so the species appears with Antilopinae. *Saiga* has been discussed above in the context of Antilopinae. The more



*Figure 6.* Supertree of Alcelaphinae, Hippotraginae and Reduncinae. Species not included: *Hippotragus leucophaeus*. The 90% majority rule tree is identical to the one shown above, except that the *Damaliscus* species appear as ((*D. hunteri*, *D. lunatus*), *D. pygargus*) in 90% of the MPTs.

speciose genera appear as ((*Capra*, *Ovis*), *Naemorhedus*), a relationship that appears to be stable.

### 3.5 Alcelaphinae, Hippotraginae and Reduncinae (Figure 6)

Evidence for relationships within these groups comes from allozymes and two mitochondrial data sources (rDNA and cytochrome *b*). These sources cluster each of the three *Damaliscus* species together and the two *Connochaetes* species together, but differ over the arrangement of these

generic groupings to one another. Data from rDNA analyzed by Essop *et al.* (1997) and Gatesy *et al.* (1997) present a largely unresolved polytomy. Allozyme data indicate *Damaliscus* and *Connochaetes* to be sister taxa (Grobler and van der Bank, 1995), whereas Matthee and Robinson (1999) give ((*Damaliscus*, *Alcelaphus*), *Connochaetes*), the topology that appears in the supertree.

*Reduncinae* appear as the sister group to the clade composed of *Alcelaphinae* and *Hippotraginae* in the supertree. The most interesting aspect of the subfamily is the placement of *Pelea capreolus*. On the basis of morphological evidence, it has been listed as having indeterminate status (Gentry, 1992), as being part of the paraphyletic tribe *Neotragini* (mainly on the basis of cranial characters; Gentry, 1990, 1992), or placed in its own subfamily (Grubb, 1993). This contrasts with the molecular evidence (rDNA, Gatesy *et al.*, 1997; cyt b, Hassanin and Douzery, 1999b), which places *Pelea* clearly as sister group to *Reduncinae*, as it appears in the supertree. *Redunca redunca* groups with each of the other *Redunca* species in the source data, but appears as the sister taxon to *R. arrundinum* in the consensus tree.

### 3.6 Cervidae (Figure 7) and Moschidae (Figure 8)

There is surprisingly little molecular information about many species of Cervidae. Groves and Grubb (1987) and Groves (2000) note the general lack of artiodactyl, and particularly deer, material in systematic collections. Only 21 of the 45 species of deer could be included here, and the composition of the data set varies: Su *et al.* (1999) provide data on musk deer and Wang and Lan (2000) provide similar information for the muntjac deer. Randi *et al.* (1998) use a sample taken across the extant cervids.

With the exception of Muntiacinae, the strict-consensus supertree for Cervidae collapses into a cone, but, as with Antilopinae, the 90% majority-rule tree is more resolved. Molecular and morphological data agree in finding a clear separation between the Odocoileinae and the Cervinae plus Muntiacinae. Differences between the two data types emerge within Odocoileinae, of which the most fundamental is placement of the antlerless *Hydropotes inermis* within Capreolinae (as a result of data from Randi *et al.*, 1998), instead of a more traditional placement as the sister group to antlered deer (Groves and Grubb, 1987; Scott and Janis, 1987). As discussed by Scott and Janis (1987), Forbes (1882) and Garrod (1877) argue that *Hydropotes* is a cervid and closely related to *Capreolus*. Brooke (1878) suggested that *Alces*, *Hydropotes*, and *Capreolus* form a natural group. This is found independently in the supertree based on data from Randi *et al.* (1998). In the

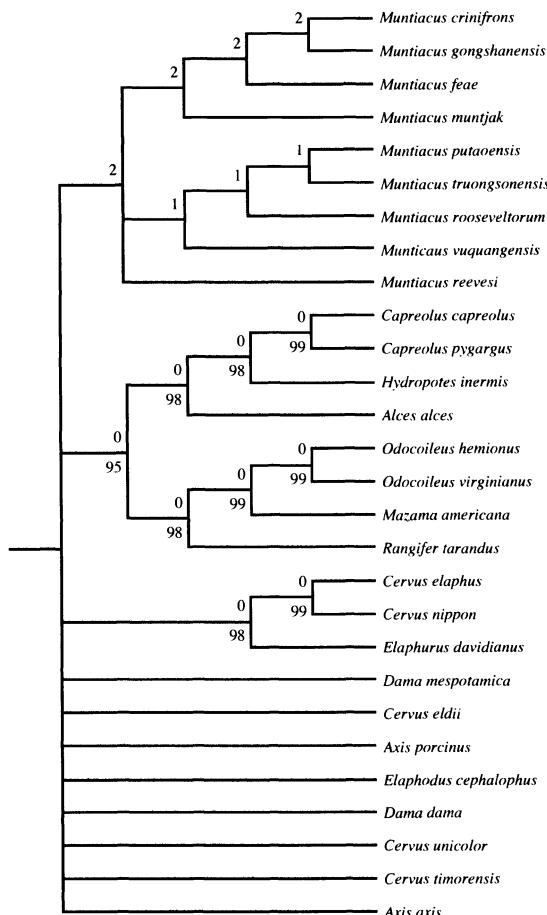


Figure 7. Supertree of Cervidae (90% majority rule consensus). Species not included: *Axis calamianensis*, *Axis kuhlii*, *Blastoceros dichotomus*, *Cervus albirostris*, *Cervus alfredi*, *Cervus mariannus*, *Hippocamelus antisensis*, *Hippocamelus bisulcus*, *Mazama bricenii*, *Mazama chunyi*, *Mazama nana*, *Mazama rufina*, *Mazama gouazoupira*, *Muntiacus atherodes*, *Ozotoceros bezoarticus*, *Pudu mephistophiles*, and *Pudu pudu*.

supertree, *Alces alces* is the sister taxon to *Capreolus* + *Hydropotes* in contrast to Groves and Grubb (1987).

In the supertree, *Cervus* is not monophyletic within Cervini, although this might be because of conflict between different taxa in the source trees. *Moschus* is excluded generally from Cervidae based on morphology. Although it has been placed close to Cervidae (Su *et al.*, 1999), it was suggested recently that it is in fact closer to Bovidae (Hassanin and Douzery, 2003), a resolution that also occurs in the supertree.

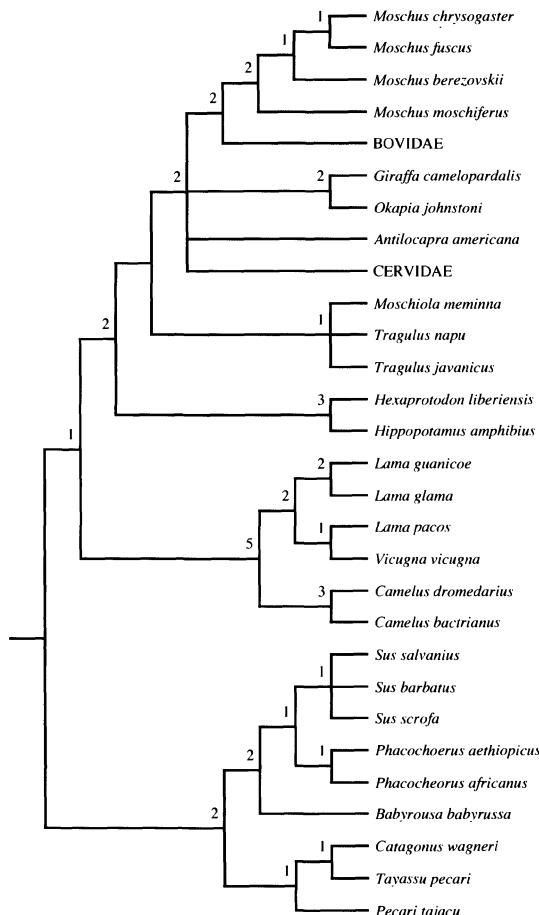


Figure 8. Supertree of Suiformes and remaining families. Species not included for Suidae: *Hyemoschus aquaticus*, *Hylochoerus meinertzhageni*, *Potamochoerus larvatus*, *Potamochoerus porcus*, *Sus bucculentus*, *Sus cebifrons*, *Sus celensis*, *Sus heureni*, *Sus philippensis*, *Sus timoriensis*, and *Sus verrucosus*.

### 3.7 Suiformes and Tragulidae (Figure 8)

Suiformes (Suidae, Tayassuidae and Hippopotamidae) are usually placed basally in artiodactyls on the basis of morphology (Gentry and Hooker, 1988). This suborder might not be monophyletic (Montgelard *et al.*, 1998). The sister-group relationship between Suidae and Tayassuidae (Suoidea; Janis *et al.*, 1998) is not disputed, but camelids have been placed between Suoidea and hippos by Matthee *et al.* (2001). They caution, however, against accepting this as definitive with the available information. In any case, a

traditional Suiformes is recovered in the supertree. Tragulidae is regarded as less advanced in many characters than the rest of the Pecora (Gentry, 1978b), and this is reflected in its basal position here.

### 3.8 Problematic taxa

The species *Aepyceros melampus*, *Saiga tatarica* and *Pelea capreolus* have proved problematic historically and have appeared in different subfamilies (Gentry, 1992). The pronghorn, *Antilocapra americana* has proven difficult also and has been placed with either Bovidae (O'Gara and Matson, 1975) or Cervidae (Gatesy *et al.*, 1992; Janis *et al.*, 1998). Here, *Antilocapra* is found with Cervidae in an unresolved polytomy. However, it is the sister taxon to Giraffidae in the 90% majority rule tree, a relationship that has been noted in the past (Beintema *et al.*, 1979).

*Saiga* and *Pelea* were discussed above in the context of Antilopinae and Reduncinae, respectively. Here, the impala, *Aepyceros*, groups with *Neotragus*, outside the main antelope clade. This relationship occurs in one of the source trees (Hassanin and Douzery, 1999b), although the authors suggest that it is a result of long-branch attraction. On morphological grounds, Neotragini are placed in the subfamily Antilopinae (Gentry, 1990), and, because only one species of *Neotragus* is included here, its position in supertree perhaps should be viewed with caution. *Aepyceros* is sometimes included within Alcelaphini (Gentry, 1978a), but is notably smaller than other members of the tribe and the females are hornless (Gentry, 1990). *Aepyceros* might share some morphological characters with Antelopini and Neotragini, but it has been thought to be an independent lineage based on its distinctive limb characters (Gentry, 1990).

## 4. Supertree / supermatrix comparison (Figure 9)

The artiodactyl supertree presented here allows for a comparison between it and the topology of Gatesy *et al.* (2002), who used a supermatrix analysis. Only species common to both analyses are shown in Figure 9; with the exception of Cetacea, the full version of the supertree contains more species than the tree derived from the supermatrix. The two trees have some similarities despite being generated from different techniques and some different source data. Suiformes, *Tragulus*, and Bovinae are identical, and Caprinae differ by one species. The supermatrix tree shows a more resolved Cervidae compared with the one derived from MRP. In addition, the two topologies differ in the arrangement of *Cephalopus*, *Gazella*, *Kobus*, and *Aepyceros*. Antilopinae and Cervidae are particularly difficult, and lack of

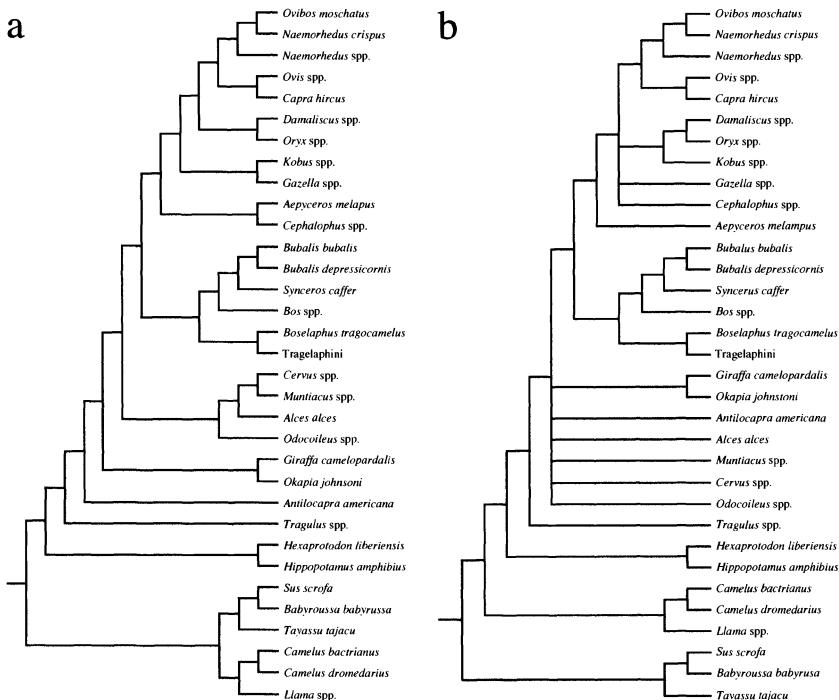


Figure 9. Comparison of the trees from a) the supermatrix study of Gatesy *et al.* (2002) and b) the supertree derived from this study pruned to the equivalent set of taxa.

data in the MRP analysis could be part of the reason for the discrepancy. Congruence between the two trees was assessed using the partition metric in Component (available at <http://taxonomy.zoology.gla.ac.uk/rod/cpw.html>). The normalized value of 0.25 indicates that there is a good level of agreement between the two trees (i.e., 75% similarity).

## 5. Conclusions

Despite some shortcomings noted, MRP remains the most accessible and tractable means of creating supertrees available currently, and has the potential to provide a comprehensive summary of phylogenetic relationships within an order of mammals. This is subject to careful selection of source trees, however. Problems remain with “jigsaw” arrangements: MRP does not always arrange fragmentary information in a satisfactory way, and suspect clades can become established. Phylogenetic study of the artiodactyls would benefit enormously from an expansion of systematic interest, especially by

molecular analysis of previously unrepresented taxa. Cervidae and antelopes are obvious areas for attention.

## Acknowledgements

I thank Adrian Friday for discussion of both the method and the artiodactyls throughout a long gestation of the supertree. I also thank Matt Symonds and Eleanor Weston who read and commented on various incarnations of the manuscript. Olaf Bininda-Emonds gave helpful advice on MRP, and I am particularly grateful to him for running both the initial ratchet searches and Bremer analysis. I thank Adrian Friday, Olaf Bininda-Emonds, Colin Groves, and two anonymous reviewers for constructive criticism of the text that helped to clarify ideas and improved it considerably. Funding from the BBSRC; Newnham College, Cambridge; The Cambridge European Trust; and The Cambridge Philosophical Society is gratefully acknowledged.

## References

References preceded by an asterisk were used as source trees (see Table 1).

- \*ALLARD, M. W., MIYAMOTO, M. M., JARECKI, L., KRAUS, F., AND TENNANT, M. R. 1992. DNA systematics and evolution of the artiodactyl family Bovidae. *Proceedings of the National Academy of Sciences of the United States of America* 89:3972–3976.
- \*AMATO, G., EGAN, M. G., AND SCHALLER, G. B. 2000. Mitochondrial DNA variation in muntjac: evidence for discovery, rediscovery and phylogenetic relationships. In E. S. Vrba and G. B. Schaller (eds), *Antelopes, Deer and Relatives: Fossil Record, Behavioural Ecology, Systematics and Conservation*, pp. 287–295. Yale University Press, New Haven and London.
- AMATO, G., EGAN, M. G., SCHALLER, G. B., BAKER, R. H., ROSENBAUM, H. C., ROBICHAUD, W. G., AND DESALLE, R. 1999. Rediscovery of Roosevelt's barking deer (*Muntiacus rooseveltorum*). *Journal of Mammalogy* 80:639–673.
- BAUM, B. R. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon* 42:3–10.
- BAUM, B. R. AND RAGAN, M. A. 2004. The MRP method. In O. R. P. Bininda-Emonds (ed). *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 17–34. Kluwer Academic, Dordrecht, the Netherlands.
- \*BEINTEMA, J. J., BREUKELMAN, H. J., DUBOIS, J.-Y. F., AND WARMELS, H. W. 2003. Phylogeny of ruminants secretory ribonuclease gene sequences of pronghorn (*Antilocapra americana*). *Molecular Phylogenetics and Evolution* 26:18–25.
- BEINTEMA, J. J., GAASTRA, W., AND MUNNIKSMA, J. 1979. Primary structure of pronghorn pancreatic ribonuclease: close relationship between giraffe and pronghorn. *Journal of Molecular Evolution* 13:305–316.
- BININDA-EMONDS, O. R. P. AND BRYANT, H. N. 1998. Properties of matrix representation with parsimony analysis. *Systematic Biology* 47:497–508.

- BININDA-EMONDS, O. R. P., GITTLEMAN, J. L., AND PURVIS, A. 1999. Building large trees by combining phylogenetic information: a complete phylogeny of the extant Carnivora (Mammalia). *Biological Reviews* 74:143–173.
- BININDA-EMONDS, O. R. P., GITTLEMAN, J. L., AND STEEL, M. A. 2002. The (super)tree of life: procedures, problems and prospects. *Annual Review of Ecology and Systematics* 33:265–289.
- BININDA-EMONDS, O. R. P., JONES, K. E., PRICE, S. A., GRENYER, M., CARDILLO, M., HABIB, A., PURVIS, A., AND GITTLEMAN, J. L. 2003. Supertrees are a necessary not-so-evil: a comment on Gatesy et al. *Systematic Biology* 52:724–729.
- BININDA-EMONDS, O. R. P., JONES, K. E., PRICE, S. A., CARDILLO, M., GRENYER, R., AND PURVIS, A. 2004. Garbage in, garbage out: data issues in supertree construction. In O. R. P. Bininda-Emonds (ed). *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 267–280. Kluwer Academic, Dordrecht, the Netherlands.
- BININDA-EMONDS, O. R. P., AND SANDERSON, M. J. 2001. Assessment of the accuracy of matrix representation with parsimony analysis supertree construction. *Systematic Biology* 50:565–579.
- \*BIRUNGI, J. AND ARCTANDER, P. 2001. Molecular systematics and phylogeny of the Reduncini (Artiodactyla: Bovidae) inferred from the analysis of mitochondrial cytochrome *b* sequences. *Journal of Mammalian Evolution* 8:125–147.
- BRASHARES, J. S., GARLAND, T. J., AND ARCESE, P. 2000. Phylogenetic analysis of coadaptation in behaviour, diet and body size in the African antelope. *Behavioural Ecology* 11:452–463.
- BREMER, K. 1988. The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. *Evolution* 42:795–803.
- BROOKE, V. 1878. On the classification of the Cervidae, with a synopsis of the existing species. *Proceedings of the zoological Society of London* 1878:883–928.
- CHIKUNI, K., MORI, Y., TABATA, T., SAITO, M., MONMA, M., AND KOSUGIYAMA, M. 1995. Molecular phylogeny based on the kappa-casein and cytochrome *b* sequences in the mammalian suborder Ruminantia. *Journal of Molecular Evolution* 41:859–866.
- COOK, C. E., WANG, Y., AND SENSABAUGH, G. 1999. A mitochondrial control region and cytochrome *b* phylogeny of sika deer (*Cervus nippon*) and report of tandem repeats in the control region. *Molecular Phylogenetics and Evolution* 12:47–56.
- \*CRONIN, M. A., STUART, R., PIERSON, B. J., AND PATTON, J. C. 1996. κ-casein gene phylogeny of higher ruminants (Pecora, Artiodactyla). *Molecular Phylogenetics and Evolution* 6:295–311.
- \*DOUZERY, E. AND RANDI, E. 1997. The mitochondrial control region of Cervidae: evolutionary patterns and phylogenetic content. *Molecular Biology and Evolution* 14:1154–1166.
- EISENBERG, J. F. 1981. *The Mammalian Radiations*. The Athlone Press, London.
- \*EMERSON, B. C. AND TATE, M. L. 1993. Genetic analysis of evolutionary relationships among deer, (subfamily Cervinae). *Journal of Heredity* 84:266–273.
- \*ESSOP, F., HARLEY, E. H., AND BAUMGARTEN, I. 1997. A molecular phylogeny of some Bovidae based on restriction-site mapping of mitochondrial DNA. *Journal of Mammalogy* 78:377–386.
- FORBES, W. A. 1882. Supplementary notes on the anatomy of the Chinese water-deer (*Hydropotes inermis*). *Proceedings of the Zoological Society of London* 636–638.
- GARROD, A. H. 1877. Notes on the anatomy of the Chinese water-deer (*Hydropotes inermis*). *Proceedings of the Zoological Society of London* 789–793.

- \*GATESY, J. 1998. Molecular evidence for the phylogenetic affinities of Cetacea. In J. G. M. Thewissen (ed.), *The Emergence of Whales*, pp. 63–111. Plenum Press, New York.
- GATESY, J., AMATO, G., VRBA, E., SCHALLER, G., AND DESALLE, R. 1997. A cladistic analysis of the mitochondrial ribosomal DNA from the Bovidae. *Molecular Phylogenetics and Evolution* 7:303–319.
- \*GATESY, J. AND ARCTANDER, P. 2000. Hidden morphological support for the phylogenetic placement of *Pseudoryx nghetinhensis* with bovine bovids: a combined analysis of gross anatomical evidence and DNA sequences from five genes. *Systematic Biology* 49:515–538.
- GATESY, J., HAYASHI, C., CRONIN, M. A., AND ARCTANDER, P. 1996. Evidence from milk casein genes that cetaceans are close relatives of hippopotamid artiodactyls. *Molecular Biology and Evolution* 13:954–963.
- GATESY, J., MATTHEE, C. A., DESALLE, R., AND HAYASHI, C. 2002. Resolution of the supertree/supermatrix paradox. *Systematic Biology* 51:652–664.
- GATESY, J. AND SPRINGER, M. S. 2004. A critique of matrix representation with parsimony supertrees. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 369–388. Kluwer Academic, Dordrecht, the Netherlands.
- GATESY, J., YELON, D., DESALLE, R., AND VRBA, E. S. 1992. Phylogeny of the Bovidae (Artiodactyla, Mammalia), based on mitochondrial ribosomal DNA sequences. *Molecular Biology and Evolution* 93:433–446.
- GENTRY, A. W. 1978a. Bovidae. In V. J. Maglio and H. B. S. Cooke (eds), *Evolution of African Mammals*, pp. 540–572. Harvard University Press, Cambridge, Massachusetts.
- GENTRY, A. W. 1978b. Tragulidae and Camelidae. In V. J. Maglio and H. B. S. Cooke (eds), *Evolution of African Mammals*, pp. 536–539. Harvard University Press, Cambridge, Massachusetts.
- GENTRY, A. W. 1990. Evolution and dispersal of African Bovidae. In G. A. Bubenik and A. B. Bubenik (eds), *Horns, Pronghorns and Antlers: Evolution, Morphology, Physiology and Social Significance*, pp. 195–227. Springer Verlag, New York.
- GENTRY, A. W. 1992. The subfamilies and tribes of the family Bovidae. *Mammal Review* 22:1–32.
- GENTRY, A. W. AND HOOKER, J. J. 1988. The phylogeny of the Artiodactyla. In M. J. Benton (ed.) *The Phylogeny and Classification of the Tetrapods*, volume 2, pp. 235–277. Clarendon Press, Oxford.
- \*GEORGIADIS, N. J., KAT, P. W., OKETCH, H., AND PATTON, J. 1990. Allozyme divergences within the Bovidae. *Evolution* 44:2135–2149.
- GINGERICH, P. D., WELLS, N. A., RUSSELL, D. E., AND SHAH, S. M. I. 1983. Origin of whales in epicontinental remnant seas: new evidence from the early Eocene of Pakistan. *Science* 220:403–406.
- GRAUR, D. AND HIGGINS, D. G. 1994. Molecular evidence for the inclusion of cetaceans within the order Artiodactyla. *Molecular Biology and Evolution* 11:357–364.
- \*GROBLER, J. P. AND VAN DER BANK, F. H. 1995. Allozyme divergence among four representatives of the subfamily Alcepaphinae (family: Bovidae). *Comparative Biochemistry and Physiology* 112:303–308.
- GROVES, C. P. 2000. Phylogenetic relationships with recent Antilopini (Bovidae). In E. S. Vrba and G. B. Schaller (eds), *Antelopes, Deer and Relatives: Fossil Record, Behavioural Ecology, Systematics and Conservation*, pp. 223–233. Yale University Press, New Haven.

- GROVES, C. P. AND GRUBB, P. 1987. Relationships of living deer. In C. M. Wemmer (ed.), *Biology and Management of the Cervidae*, pp. 21–59. Smithsonian Institution Press, Washington. *The Symposia of the National Zoological Park*.
- \*GROVES, C.P. AND SHIELDS, G. F. 1996. Phylogenetics of the Caprinae based on cytochrome *b* sequence. *Molecular Phylogenetics and Evolution* 5:467–476.
- \*GRUBB, P. 1993. Order Artiodactyla. In D. E. Wilson and D. M. Reeder (eds), *Mammal Species of the World: A Taxonomic and Geographic Reference*, pp. 377–414. Smithsonian Institution Press, Washington.
- \*HARTL, G. B., BURGER, H., WILLING, R. AND SUCHENTRUNK, F. 1990. On the biochemical systematics of the Caprini and Rupicaprini. *Biochemical Systematics and Ecology* 18:175–182.
- \*HASSANIN, A. AND DOUZERY, E. J. P. 1999a. Evolutionary affinities of the enigmatic saola (*Pseudoryx nghetinhensis*) in the context of the molecular phylogeny of Bovidae. *Proceedings of the Royal Society of London Series B, Biological Sciences* 266:893–900.
- \*HASSANIN, A. AND DOUZERY, E. J. P. 1999b. The tribal radiation of the family Bovidae (Artiodactyla) and the evolution of the mitochondrial cytochrome *b* gene. *Molecular Phylogenetics and Evolution* 13:227–243.
- \*HASSANIN, A. AND DOUZERY, E. J. P. 2003. Molecular and morphological phylogenies of the Ruminantia and the alternative position of the Moschidae. *Systematic Biology* 52:206–228.
- \*HASSANIN, A., PASQUET, E., AND VIGNE, J.-D. 1998. Molecular systematics of the subfamily Caprinae (Artiodactyla, Bovidae) as determined from cytochrome *b* sequences. *Journal of Mammalian Evolution* 5:217–236.
- IRWIN, D. M. AND ARNASON, U. 1994. Cytochrome *b* gene of marine mammals: phylogeny and evolution. *Journal of Mammalian Evolution* 2:37–55.
- IRWIN, D. M., KOCHER, T. D., AND WILSON, A. C. 1991. Evolution of the cytochrome *b* gene of mammals. *Journal of Molecular Evolution* 32:128–144.
- \*JANECEK, L. L., HONEYCUTT, R. L., ADKINS, R. M., AND DAVIS, S. K. 1996. Mitochondrial gene sequences and the molecular systematics of the artiodactyl subfamily Bovinae. *Molecular Phylogenetics and Evolution* 6:107–119.
- JANIS, C. M., EFFINGER, J. A., HARRISON, J. A., HONEY, J. G., KRON, D. G., LANDER, B., MANNING, E., PROTHERO, D. R., STEVENS, M. S., STUCKY, R. K., WEBB, S. D., AND WRIGHT, D. B. 1998. Artiodactyla. In C. M. Janis, K. M. Scott and L. L. Jacobs (eds), *Tertiary Mammals of North America*, volume 1: Terrestrial Carnivores, Ungulates and Ungulatelike Mammals, pp. 337–357. Cambridge University Press, Cambridge.
- JONES, K. E., PURVIS, A., MACLARNON, A., BININDA-EMMONDS, O. R. P., AND SIMMONS, N. B. 2002. A phylogenetic supertree of the bats (Mammalia: Chiroptera). *Biological Reviews* 77:223–259.
- \*KLEINEIDAM, R. G., PESOLE, G., BREUKELMAN, H. J., BEINTEMA, J. J., AND KASTELEIN, R. A. 1999. Inclusion of cetaceans within the order Artiodactyla based on phylogenetic analysis of pancreatic ribonuclease genes. *Journal of Molecular Evolution* 48:360–368.
- LIU, F.-G. R., MIYAMOTO, M. M., FREIRE, N. P., ONG, P. Q., TENNANT, M. R., YOUNG, T. S., AND GUGEL, K. F. 2001. Molecular and morphological supertrees for eutherian (placental) mammals. *Science* 291:1786–1789.
- LOWE, V. P. W. AND GARDINER, A. S. 1989. Are the new and old world wapitis (*Cervus canadensis*) conspecific with red deer (*Cervus elaphus*)? *Journal of Zoology, London* 218:51–58.
- \*MADSEN, O., SCALLY, M., DOUADY, C., KAO, D. J., DERBY, R. W., ADKINS, R. M., AMRINE, H. M., STANHOPE, M. J., DE JONG, W. W., AND SPRINGER, M. S. 2001. Parallel adaptive radiations in two major clades of placental mammals. *Nature* 409:610–614.

- \*MATTHEE, C. A., BURZLAFF, J. D., TAYLOR, J. F., AND DAVIS, S. K. 2001. Mining the mammalian genome for artiodactyl systematics. *Systematic Biology* 50:367–390.
- \*MATTHEE, C. A. AND ROBINSON, T. J. 1999. Cytochrome *b* phylogeny of the family Bovidae: resolution within the Alcepini, Antilopini, Neotragini and Tragelaphini. *Molecular Phylogenetics and Evolution* 12:31–46.
- \*MIYAMOTO, M. M., KRAUS, F., AND RYDER, O. A. 1990. Phylogeny and evolution of antlered deer determined from mitochondrial DNA sequences. *Proceedings of the National Academy of Sciences of the United States of America* 87:6127–6131.
- \*MIYAMOTO, M. M., TANHAUSER, S. M., AND LAIPIIS, P. 1989. Systematic relationships in the artiodactyl tribe Bovini (family Bovidae), as determined from mitochondrial DNA sequences. *Systematic Zoology* 38:342–349.
- MOLINA, M. AND MOLINARI, J. 1998. Taxonomy of Venezuelan white-tailed deer (*Odocoileus*, Cervidae, Mammalia), based on cranial and mandibular traits. *Canadian Journal of Zoology* 77:632–645.
- \*MONTGELARD, C., DUCROCQ, S., AND DOUZERY, E. 1998. What is a suiforme (Artiodactyla)? Contribution of cranioskeletal and mitochondrial DNA data. *Molecular Phylogenetics and Evolution* 9:528–532.
- \*MURPHY, W. J., EIZIRIK, E., JOHNSON, W. E., ZHANG, Y. P., RYDER, O. A., AND O'BRIEN, S. J. 2001. Molecular phylogenetics and the origins of placental mammals. *Nature* 409:614–618.
- \*NIKAIDO, M., ROONEY, A. P., AND OKADA, N. 1999. Phylogenetic relationships among cetartiodactyls based on insertions of short and long interspersed elements: hippopotamuses are the closest extant relatives of whales. *Proceedings of the National Academy of Sciences of the United States of America USA* 96:10261–10266.
- NIXON, K. C. 1999. The parsimony ratchet, a new method for rapid parsimony analysis. *Cladistics* 15:407–414.
- O'GARA, B. W. AND MATSON, G. 1975. Growth and casting of horns by pronghorns and exfoliation of horns by bovids. *Journal of Mammalogy* 56:829–846.
- PAGE, R. D. M. 1996. TreeView: an application to display phylogenetic trees on personal computers. *Computer Applications in the Biosciences* 12:357–358.
- PISANI, D., YATES, A. M., LANGER, M. C., AND BENTON, M. J. 2002. A genus-level supertree of the Dinosauria. *Proceedings of the Royal Society of London Series B, Biological Sciences* 269:915–921.
- \*PITRA, C., FÜRBASS, R., AND SEYFERT, H.-M. 1997. Molecular phylogeny of the tribe Bovini (Mammalia: Artiodactyla): alternative placement of the Anoa. *Journal of Evolutionary Biology* 10:589–600.
- \*POLZIEHN, R. O. AND STROBECK, C. 1998. Phylogeny of wapiti, red deer, sika deer and other North American cervids as determined from mitochondrial DNA. *Molecular Phylogenetics and Evolution* 10:249–258.
- PROTHERO, D. R. 1993. Ungulate phylogeny: molecular vs. morphological evidence. In F. Szalay, M. J. Novacek and M. C. McKenna (eds), *Mammal Phylogeny: Placentals*, pp. 173–181. Springer-Verlag, New York.
- PURVIS, A. 1995. A composite estimate of primate phylogeny. *Philosophical Transactions of the Royal Society London B* 348:405–421.
- PURVIS, A. AND WEBSTER, A. J. 1999. Phylogenetically independent comparisons and primate phylogeny. In P. Lee (ed.), *Comparative Primate Socioecology*, pp. 44–69. Cambridge University Press, Cambridge.
- RAGAN, M. A. 1992. Phylogenetic inference based on matrix representation of trees. *Molecular Phylogenetics and Evolution* 1:53–58.

- \*RANDI, E., FUSCO, G., R., L., TOSCO, S., AND TOSI, G. 1991. Allozyme divergence and phylogenetic relationships among *Capra*, *Ovis* and *Rupicapra* (Artiodactyla, Bovidae). *Heredity* 67:281–286.
- \*RANDI, E., LUCCHINI, V., AND DIONG, C. H. 1996. Evolutionary genetics of the suiformes as reconstructed using mtDNA sequencing. *Journal of Mammalian Evolution* 3:163–194.
- \*RANDI, E., MUCCI, N., PIERPAOLI, M., AND DOUZERY, E. 1998. New phylogenetic perspectives on the Cervidae (Artiodactyla) are provided by the mitochondrial cytochrome *b* gene. *Proceedings of the Royal Society of London Series B, Biological Sciences* 265:793–801.
- ROBINSON, T. J., HARRISON, W. R., DE LEÓN, F. A. P., DAVIS, S. K., AND ELDER, F. F. B. 1998. A molecular cytogenetic analysis of X chromosome repatterning in the Bovidae: transposition, inversion and phylogenetic inference. *Cytogenetics and Cell Genetics* 80:179–184.
- ROBINSON, T. J., WILSON, V., GALLAGHER JR., D. S., TAYLOR, J. F., DAVIS, S. K., HARRISON, W. R., AND ELDER, F. F. B. 1996. Chromosomal evolution in duiker antelope (Cephalopinae: Bovidae): karyotype comparisons, fluorescence *in situ* hybridization and rampant X chromosome variation. *Cytogenetics and Cell Genetics* 73:116–122.
- SÆTHER, B.-E. AND GORDON, I. J. 1994. The adaptive significance of reproductive strategies in ungulates. *Proceedings of the Royal Society of London Series B, Biological Sciences* 256:263–268.
- SANDERSON, M. J., PURVIS, A., AND HENZE, C. 1998. Phylogenetic supertrees: assembling the trees of life. *Trends in Ecology and Evolution* 13:105–109.
- SCHALLER, G. B. AND VRBA, E. S. 1996. Description of the giant muntjac (*Megamuntiacus vuquangensis*) in Laos. *Journal of Mammalogy* 77:675–683.
- \*SCHREIBER, A., SEIBOLD, I., NOTZOLD, G., AND WINK, M. 1999. Cytochrome *b* gene haplotypes characterize chromosomal lineages of Anoa, the Sulawesi dwarf buffalo. *Journal of Heredity* 90:165–176.
- SCOTT, K. M. AND JANIS, C. M. 1987. Phylogenetic relationships of the Cervidae and the case for a superfamily “Cervoidea”. In C. M. Wemmer (ed.), *Biology and Management of the Cervidae*, pp. 3–20. Smithsonian Institution Press, Washington.
- SHIMAMURA, M., YASUE, H., OHSHIMA, K., ABE, H., KATO, H., KISHIRO, T., GOTO, M., MUNECHIKA, I., AND OKADA, N. 1997. Molecular evidence from retroposons that whales form a clade within even-toed ungulates. *Nature* 388:666–670.
- SHOSHANI, J. AND MCKENNA, M. 1998. Higher taxonomic relationships among extant mammals based on morphology, with selected comparisons of results from molecular data. *Molecular Phylogenetics and Evolution* 9:572–584.
- SMITH, M. R., SHIVJI, M. S., WADDELL, V. G., AND STANHOPE, M. J. 1996. Phylogenetic evidence from the IRBP genes for the paraphyly of toothed whales, with mixed support for Cetacea as a suborder of Artiodactyla. *Molecular Biology and Evolution* 13:918–922.
- SPRINGER, M. S. AND DE JONG, W. W. 2001. Which mammalian supertree to bark up? *Science* 291:1709–1711.
- \*STANLEY, H. F., KADWELL, M., AND WHEELER, J. C. 1994. Molecular evolution of the family Camelidae: a mitochondrial DNA study. *Proceedings of the Royal Society of London Series B, Biological Sciences* 256:1–6.
- \*SU, B., WANG, Y.-X., LAN, H., WANG, W., AND ZHANG, Y. 1999. Phylogenetic study of complete cytochrome *b* genes in musk deer (genus *Moschus*) using museum samples. *Molecular Phylogenetics and Evolution* 12:241–249.
- SWOFFORD, D. L. 2002. *PAUP\**. *Phylogenetic Analysis Using Parsimony (\*and Other Methods)*. Version 4. Sinauer, Sunderland, Massachusetts.

- \*THEIMER, T. C. AND KEIM, P. 1998. Phylogenetic relationships of peccaries based on mitochondrial cytochrome *b* DNA sequences. *Journal of Mammalogy* 79:566–572.
- \*VAN VIUREN, B. J. AND ROBINSON, T. J. 2001. Retrieval of four adaptive lineages in duiker antelopes: evidence from mitochondrial DNA sequences and fluorescence *in situ* hybridization. *Molecular Phylogenetics and Evolution* 20:409–425.
- \*VASSART, M., SÉGUÉLA, A., AND HAYES, H. 1995. Chromosomal evolution in gazelles. *Journal of Heredity* 86:216–227.
- VÁZQUEZ, D. P. AND GITTLEMAN, J. L. 1998. Biodiversity conservation: does phylogeny matter? *Current Biology* 8:R379–R381.
- \*WALL, D. A., DAVIS, S. K. AND READ, B. M. 1992. Phylogenetic relationships in the subfamily Bovinae (Mammalia: Artiodactyla) based on ribosomal DNA. *Journal of Mammalogy* 73:262–275.
- \*WANG, W. AND LAN, H. 2000. Rapid and parallel chromosomal number reductions in muntjac deer inferred from mitochondrial DNA phylogeny. *Molecular Biology and Evolution* 17:1326–1333.
- WILSON, D. E. AND REEDER, D. M. 1993. *Mammal Species of the World: A Taxonomic and Geographic Reference*. Smithsonian Institution Press, Washington.

# Chapter 20

## SUPERTREES

*Using complete phylogenies in comparative biology*

John L. Gittleman, Kate E. Jones, and Samantha A. Price

**Abstract:** The field of comparative biology has undergone a renaissance since researchers began examining macroevolutionary questions in an explicit phylogenetic framework. However, research into these new areas is now hampered by incomplete phylogenetic information for the clades of interest. Species-level trees made by supertree construction techniques offer a way to generate complete phylogenies. Here we review the types of questions that can now be addressed and evaluate critically when using complete trees made by supertree construction techniques might be inappropriate.

**Keywords:** biodiversity; comparative methods; conservation; extinction; phylogenies; speciation; trait evolution

### 1. Introduction

Since Darwin (1859), the comparative method has been used to ask questions about patterns of evolutionary change. It has long been recognized that there are two components affecting patterns of evolutionary change: that of adaptation to environments and of lineage-specific effects that are invariant with the environment (Darwin's "conditions of existence" and "unity of type", respectively). However, only recently have more formalized techniques been developed to measure the relative contributions of these two components by incorporating the pattern of evolutionary change directly into comparative analyses, thereby leading to a re-invigoration of this field (Harvey and Pagel, 1991). Application of these comparative phylogenetic techniques have led to key insights into a diverse array of important biological questions over the past decade, including patterns of correlated

trait evolution (e.g., evidence for the mosaic evolution of the mammalian brain; Barton and Harvey, 2000) and patterns of co-speciation (e.g., co-evolution of pocket gophers and their parasitic lice; Morand *et al.*, 2000). Complete phylogenies are necessary for answering questions about the evolutionary process itself — speciation and extinction — and for controlling lineage-specific effects when studying patterns of adaptation.

Unfortunately, complete trees are difficult to generate. Molecular sampling in different clades is often opportunistic, leading to poor clade coverage. Sampling for morphological traits is usually more complete, but trees generated by these data are often less formal analytically and of poor resolution (e.g., taxonomies). Supertree construction techniques offer exciting new opportunities to examine the nature of evolutionary processes because they can generate complete phylogenies of entire large clades quickly based on multiple lines of evidence (Bininda-Emonds *et al.*, 2002). Additionally, supertrees introduce less error as a result of taxon sampling artifacts, and therefore provide a means for truly examining macrobiological patterns across complete clades. In this chapter, we consider what kind of evolutionary questions one can ask uniquely with complete phylogenies and the possible future applications of such phylogenies. Additionally, we evaluate critically when using complete trees made by supertree construction is inappropriate to examine these new questions. We focus here on comparative evolutionary applications for complete supertrees rather than for trees in general because the latter is covered elsewhere (e.g., Harvey and Pagel, 1991).

## 2. Descriptive systematics and priority setting

Most of our discussion here deals with unique uses of supertrees for comparative hypothesis testing because supertrees allow us to gain access to comprehensive and large phylogenies. At the outset, therefore, we want to discuss briefly how supertrees can be used to diagnose whether taxonomic completeness is possible. Complete phylogenies made by supertree construction techniques are an extremely valuable tool in descriptive systematics. In the process of culling the phylogenetic information for all species within a clade from published phylogenies, differences in systematic effort among taxa can be quantified, thereby identifying groups that are desperately in need of more research and providing a starting point for future studies. For example, a recent supertree study of Chiroptera (bats) revealed that over one-third of all phylogenetic studies have investigated one family (Phyllostomidae), although it represents only one-sixth of all bat species (Jones *et al.*, 2002). At the other extreme, several bat clades have never been

**Table 1.** Phylogenetic resolution and sampling effort in different bat clades (OW = Old World, NW = New World). For each clade,  $N_{\text{taxa}}$  = number of taxa; %<sub>RES</sub> = resolution of the supertree topology as a percentage of a fully bifurcating solution;  $N_{\text{sour}}$  = number of independent source trees; and  $N_{\text{char}}$  = number of binary characters recoded from the source tree topologies into the supertree matrix (Jones *et al.*, 2002). Poor resolution results from both poor coverage per species ( $N_{\text{char}} / N_{\text{taxa}}$ ) and poor sampling in each study ( $N_{\text{char}} / N_{\text{sour}} / N_{\text{taxa}}$ ) for Kerivoulinae and Rhinolophidae. However, low resolution in Megadermatidae, Hipposideridae and Pteropodidae is more likely a result of disagreements between source trees as the coverage per species and sampling level for each clade is relatively high.

| Clade                                  | $N_{\text{taxa}}$ | % <sub>RES</sub> | $N_{\text{sour}}$ | $N_{\text{char}}$ | $N_{\text{char}} / N_{\text{taxa}}$ | $N_{\text{char}} / N_{\text{sour}} / N_{\text{taxa}}$ |
|----------------------------------------|-------------------|------------------|-------------------|-------------------|-------------------------------------|-------------------------------------------------------|
| Kerivoulinae<br>(Wooly bats)           | 22                | 10.0             | 2                 | 4                 | 0.18                                | 0.09                                                  |
| Rhinolophidae<br>(Horseshoe bats)      | 64                | 17.7             | 5                 | 58                | 0.91                                | 0.18                                                  |
| Megadermatidae<br>(False vampire bats) | 5                 | 33.3             | 4                 | 8                 | 1.60                                | 0.40                                                  |
| Vespertilioninae<br>(Vesper bats)      | 268               | 35.6             | 32                | 321               | 1.20                                | 0.04                                                  |
| Murininae<br>(Tube-nosed bats)         | 16                | 35.7             | 2                 | 8                 | 0.50                                | 0.25                                                  |
| Hipposideridae<br>(OW leaf-nosed bats) | 66                | 35.9             | 6                 | 85                | 1.29                                | 0.21                                                  |
| Pteropodidae<br>(OW fruit bats)        | 166               | 46.1             | 14                | 265               | 1.60                                | 0.11                                                  |
| Miniopterinae<br>(Long-fingered bats)  | 10                | 50.0             | 2                 | 5                 | 0.50                                | 0.25                                                  |
| Molossidae<br>(Free-tailed bats)       | 80                | 56.4             | 12                | 125               | 1.56                                | 0.13                                                  |
| Nycteridae<br>(Slit-faced bats)        | 12                | 60.0             | 3                 | 14                | 1.17                                | 0.39                                                  |
| Phyllostomidae<br>(NW leaf-nosed bats) | 141               | 66.2             | 39                | 630               | 4.47                                | 0.11                                                  |
| Natalidae<br>(Funnel-eared bats)       | 5                 | 66.7             | 2                 | 4                 | 0.80                                | 0.40                                                  |
| Emballonuridae<br>(Sheath-tailed bats) | 47                | 68.9             | 11                | 113               | 2.40                                | 0.22                                                  |
| Mormoopidae<br>(Naked-backed bats)     | 8                 | 83.3             | 7                 | 23                | 2.88                                | 0.41                                                  |

investigated cladistically (Kerivoulinae, Miniopterinae, Murininae, Natalidae, and Rhinopomatidae; Table 1), and should be a natural focus for future research. The disproportionate information available for certain taxa seems a function of which groups are viewed as charismatic and economically valued, and little to do with their importance to the diversity of the clade. For example, a supertree study of all eutherian (placental)

mammals found only ten clades that were represented in more than 45% of the total information that was used to construct the supertree: Bovidae (cows), Balaenopteridae and Delphinidae (whales and dolphins), Carnivora (dogs and cats), Caviidae and Muridae (cavies and rats), Equidae (horses), Leporidae (rabbits and hares), Primates, and Suidae (pigs) (Liu *et al.*, 2001). The analyses implemented when building supertrees are useful in identifying such trends.

The relative support for relationships of taxa within different clades can also be quantified using this approach and used to set priorities for future research by distinguishing between those relationships that are poorly resolved because they are less studied or because of disagreement among the source trees (Table 1). Complete trees based on supertree techniques can also be used to assess differences among phylogenies for a given group through a “sliding window” time-series analysis that shows how species have fit into published phylogenies over time. For example, such an analysis from supertree data collection of phylogenies from 1869 to 1999 showed that the giant panda was always held unequivocally to be more closely related to bears than to raccoons, even before the advent of molecular analysis (Bininda-Emonds, *in press*).

### 3. Trait co-evolution

Traditionally, comparative biology was interested in how traits co-evolve with one other. For example, for the relationship between diet and brain size in primates, it has been shown that fruit-eating species have larger neocortex sizes relative to the size of the rest of their brains (Barton, 1998). However, it was recognized that such comparisons might violate statistical assumptions when the methods assume that data points are independent of one other (Felsenstein, 1985; Harvey and Pagel, 1991). Because species are related through evolutionary descent, they often inherit similar traits; therefore, it is unlikely that the trait values of species are independent statistically (although not impossible under different trait evolutionary models; Losos, 2000; Martins *et al.*, 2002; see below). Phylogeny is now incorporated routinely in comparative biology to control for this statistical non-independence. The use of complete phylogenies has helped expand the scope of the questions that can be addressed across diverse, independent clades: for example, in studies of convergent evolution in vascular plants (Linder, 2000), abundance patterns in Australian marsupials (Johnson, 1998), immune system functions in primates and carnivores (Nunn *et al.*, 2000, 2003), and correlates of extinction risk in primates, carnivores, and bats (Purvis *et al.*, 2000b; Jones *et al.*, 2003). Additionally, questions about the rate and nature of trait change

*Table 2.* Phylogenetic signal in life history and ecology traits in mammals.  $N$  = sample size,  $\lambda$  = the maximum likelihood estimate of the phylogenetic signal parameter for each trait (values of 0 indicate traits have no phylogenetic component, whereas values of 1 indicate that the trait is correlated perfectly with phylogeny), ln lik = log likelihood of the model when  $\lambda = 0$  and significance of values of  $\lambda$  being greater than 0. \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$  (adapted from Freckleton *et al.*, 2002).

| Trait           | $N$ | $\lambda$ | ln lik, $\lambda = 0$ |
|-----------------|-----|-----------|-----------------------|
| Body mass       | 60  | 1.00      | -161.31 ***           |
| Diet            | 60  | 1.00      | -95.21 ***            |
| Latitude        | 60  | 0.67      | -230.68 **            |
| Home range size | 43  | 0.00      | -91.25                |
| Metabolic rate  | 26  | 1.00      | -54.17 ***            |
| Lifespan        | 26  | 1.00      | -27.25 ***            |

across the phylogeny can now be examined more comprehensively by using complete phylogenies that have some measure of divergence in units of time (branch lengths). Dating techniques are in their infancy for complete trees where not all taxa have some measure of sequence divergence (Purvis, 1995; Bininda-Emonds *et al.*, 1999), but are improving rapidly (Sanderson, 2003; Bryant, 2004; Vos and Mooers, 2004; Jones *et al.*, in prep.). Here, we review three areas where complete and large supertrees will improve analyses of trait co-evolution: quantifying phylogenetic signal in traits, examining models of trait evolution, and identifying sister taxa to generate statistically matched pairs.

### 3.1 Phylogenetic signal in comparative biology

Recent interest has focused on quantifying how much traits are influenced by phylogeny or contain “phylogenetic signal” (Freckleton *et al.*, 2002; Blomberg *et al.*, 2003; see Table 2). It has become important to quantify how much phylogenetic signal each trait shows because the newer, more powerful phylogenetic comparative methods control for the exact amount of phylogenetic signal in each trait when investigating trait co-evolution (e.g., Pagel, 1999; Blomberg *et al.*, 2003). There are now several different methods of estimating the amount of phylogenetic signal in traits (e.g., Cheverud *et al.*, 1985; Gittleman and Kot, 1990; Abouheif, 1999; Pagel, 1999; Blomberg *et al.*, 2003). These tests are dependent on the clade size and the accuracy of the tree topology and trait data (Bininda-Emonds and Gittleman, 2000; Blomberg *et al.*, 2003). Using complete phylogenetic information about the clade can improve the accuracy of both the topology (see Hillis, 1996) by reducing the effects of long-branch attraction and the

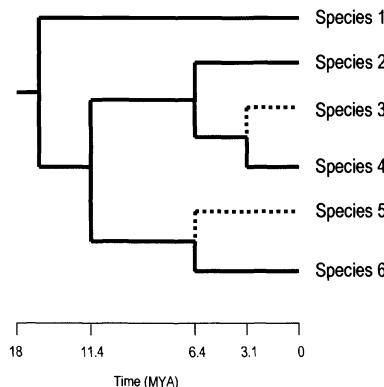
measure of phylogenetic signal in the traits by increasing the number of taxa in each clade that can be considered.

### 3.2 Models of trait evolution

By using more complete, large supertrees, it is also possible to test models of how different traits have evolved. In most phylogenetic comparative methods, models of trait change are assumed in which closely related species have traits that are more similar than those between more distantly related ones (Felsenstein, 1985). Two common models assume that trait change is directly proportional to either elapsed time or to the number of observed speciation events (referred to typically as the “Brownian motion” and “punctuational” models, respectively; see Harvey and Pagel, 1991; Purvis *et al.*, 1994). However, the accuracy of these models is unknown empirically, although each is an effective null model for various phylogenetic comparative tests (Purvis *et al.*, 1994) and hundreds of comparative papers hinge on them. Using other models of trait evolution, of which there are many (see Harvey and Rambaut, 2000), closely related taxa can have traits that are more different than between distantly related taxa such that the amount of trait evolution does not scale to the phylogeny (Price, 1997; Losos, 2000; Martins *et al.*, 2002). With complete supertrees that include divergence times, characters can be examined across the tree to examine how traits are changing relative to what would be expected from each model. Simulation studies indicate clearly that these non-Brownian-motion models are viable theoretically (Harvey and Rambaut, 2000), and complete trees can tell us which traits follow which models, with revisions of comparative statistical methods to follow suit.

### 3.3 Sister taxa comparisons

In comparative biology, comparing sister taxa is an extremely powerful tool to examine patterns of trait evolution. Sister taxa are the same age (they diverged at the same time from a common ancestor) and share many traits, and so form a statistically matched pair. Complete trees are crucial to identify sister taxa because the identified closest related taxon could be incorrect if the phylogenetic information is incomplete (Figure 1). Using complete phylogenies enables accurate identification of matched pairs on which to test hypotheses, and this is particularly important when the question relies on an estimate of time since divergence of these two taxa. We review briefly three areas of research that rely typically on comparisons across sister taxa and would benefit particularly from complete phylogenies.



*Figure 1.* Estimating sister taxa and divergence times. Missing taxa (dashed lines) can cause sister taxa to be estimated incorrectly (e.g., species 2 would be considered wrongly to be the sister taxon of species 4) and divergence times to be overestimated (e.g., divergence time for species 4 from its most recent common ancestor would be calculated as 6.4 million years, not 3.1).

### 3.3.1 Age and area models

Interest in geographic range evolution has focused historically on how range size changes over the evolutionary lifespan of a taxon: at speciation, a new species inherits a proportion of the range of its ancestor and, at extinction, range size declines to zero. Several models of range-size change have been proposed (see Gaston, 1998; Jones *et al.*, in press for reviews), but diagnosis of these models requires that the range of a species is known throughout its evolutionary history. An alternative approach is to examine interspecific variation in range sizes of contemporary species as representative of an intraspecific trend (e.g., Webb and Gaston, 2000). Here, phylogenetic age (the age at which a species diverged from its sister taxa; see Figure 1) is correlated against current range size. Measuring phylogenetic age in this manner requires that the phylogenetic information for all the taxa in the clade of interest is known. Missing taxa would overestimate the phylogenetic age of all species in the phylogeny (extinct taxa could also have the same effect, although obviously this is more difficult to account for). Using complete trees is crucial to estimate phylogenetic age correctly and to investigate these questions.

### 3.3.2 Speciation models

Sister-pair comparisons have been used to identify the spatial mode of speciation in different clades by investigating the overlap of their geographic

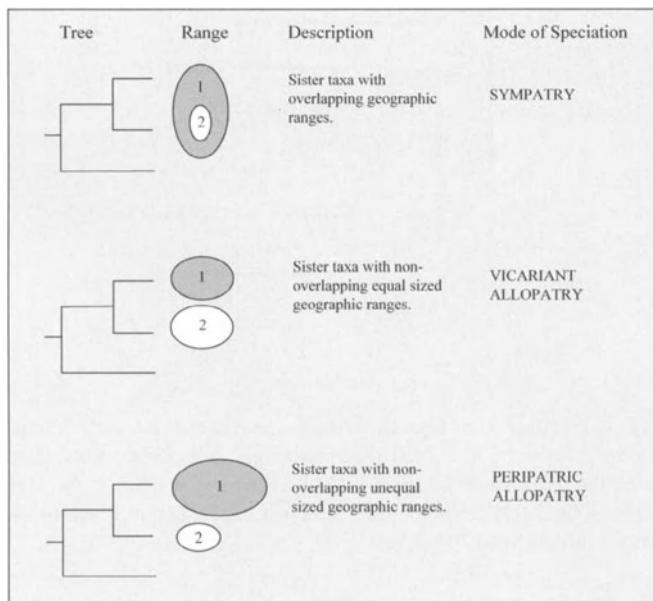


Figure 2. Models of speciation. Current geographic distributions of sister taxa (species 1 and 2) can be used to infer modes of speciation (adapted from Losos and Glor, 2003).

range distributions (Figure 2). With complete trees, the pattern of geographic ranges can be examined with reference to the entire phylogeny (Barraclough *et al.*, 1998; Barraclough and Nee, 2001; Losos and Glor, 2003). The degree of sympatry (overlap) in geographic ranges of sister taxa can be estimated and plotted against their time since divergence. The mode of speciation (e.g., allopatric or sympatric) can be determined from the deviations from a null model (e.g., geographic range shifts are stochastic following speciation events). Clearly, if the phylogeny is incomplete (or there are “hidden extinctions”), then the estimated degree of overlap and time since divergence between sister taxa can be wrong. Many studies suggest that an allopatric pattern is the most common mode of speciation (e.g., in primates, fruit flies, and North American tiger beetles; Barraclough *et al.*, 1998). It is hoped that further analyses with more complete species-level trees will test the generality of these patterns in a broader range of taxa. As Losos and Glor (2003) point out, the assumptions that these methods make (e.g., assuming that current ranges reflect the distribution of species at speciation) need to be tested more rigorously before these results are convincing.

The correlation between phenotypic disparity and the mode of speciation can also be examined using sister-pair comparisons. For example, there

might be a positive relationship between phenotypic dissimilarity and degree of sympatry if phenotypic differences are necessary for sympatric species to co-exist. Alternatively, there might be a negative relationship if phenotypic variation evolves as a consequence of geographical variation in environmental conditions. There is some evidence that different patterns operate in similar clades (e.g., tiger beetles; Barraclough *et al.*, 1998; Barraclough and Vogler, 2000), and further tests are needed across more complete phylogenies.

### 3.3.3 Trait evolutionary lags

One of the most common reasons given for failure to find significant correlation among traits is the “evolutionary lag” phenomenon (see Harvey and Pagel, 1991). A well-cited example of this phenomenon is for brain:body size allometry in mammals. Observed slopes of brain size on body size are very often less than isometric, with taxonomic differences revealing shallower slopes at higher levels, suggesting evolutionary lag of brain size not responding to selection as quickly as body size (Lande, 1979; Pagel and Harvey, 1988). Deaner and Nunn (1999) developed a method for measuring relative change of one quantitative trait compared with another (e.g., brain versus body size) along the branches of a phylogeny without assuming any explicit model of evolution (but requiring meaningful branch lengths). As one trait changes along a branch, it is measured relative to the other trait changing along the same branch; comparisons are then made for all sister taxa. Using this method on the complete species-level primate supertree (Purvis, 1995), no evidence was found for evolutionary lag in brain size. As with all analyses that rely on phylogenetic age, it is crucial that all taxa in the clade are present; otherwise, an observed pattern might result from missing taxa rather than being a true pattern resulting from all trait values. With complete sampling, the approach can be used to examine the rate of evolution of different traits more widely (e.g., Purvis *et al.*, 2003; see also Moore *et al.*, 2004). Without a supertree, it would not be possible to examine rate change between traits because a phylogeny must be fairly large and comprehensive, characteristics that are rare even for well-studied groups.

## 4. Patterns and processes of cladogenesis

Considerable attention in comparative biology is now turning to patterns and processes of cladogenesis (Allen *et al.*, 2002). A glance at almost every phylogeny reveals that some clades have more species than others. However,

to examine clade-richness meaningfully, comparisons of clade size need to be made between clades of the same age (Barracough *et al.*, 1998). This can be done by comparing the sizes of sister clades (by definition the same age) and / or by estimating divergence times across the entire tree to determine rates of cladogenesis. For either analysis, complete trees for the clade of interest are crucial (Barracough *et al.*, 1998), and provide us with an exciting opportunity to test hypotheses about the evolutionary process itself. Here, we discuss developments in two current areas: 1) whether rates of cladogenesis differ between clades and 2) identifying biological or ecological correlates of different cladogenesis rates.

## 4.1 Rates of cladogenesis

Investigation into rates of cladogenesis relies either on the topology of the tree to compare diversification rates of sister clades (e.g., Slowinski and Guyer, 1989; Kirkpatrick and Slatkin, 1993; Moore *et al.*, 2004) or on the branch lengths (nodal dates) of the tree (e.g., Nee *et al.*, 1992, 1994) to estimate rates of cladogenesis. For methods that rely on the topology alone, a complete tree is necessary so that sister-taxa can be chosen accurately. Additionally, to obtain a measure of tree balance (Kirkpatrick and Slatkin, 1993), it is imperative to have all the species within a clade represented; otherwise, the comparisons between the number of taxa in each clade will be affected by taxon sampling. For the methods that use nodal dates within birth-death or coalescent processes to estimate rates of cladogenesis (e.g., Nee *et al.*, 1992, 1994), a complete phylogeny is not required. Methods have been developed to take missing taxa into account on trees built from molecular sequences (Pybus and Harvey, 2000; Pybus *et al.*, 2002). However, if a molecular tree is not available, it is important to have a complete tree to allow accurate calculation of branch lengths (see Section 7.1). Both topology and branch-length based methods have been used in conjunction with supertrees to look at evolutionary processes. For example, a constant rates birth-death null model (Nee *et al.*, 1992, 1994) was used to identify clades that contain more species than expected in the complete species-level supertree of the primates (Purvis, 1995). Cercopithecidae (Old World monkeys) were found to have a higher diversification rate than other primate lineages (Purvis *et al.*, 1995), a result that was found later to hold irrespective of null model used (Paradis, 1998; see also Moore *et al.*, 2004).

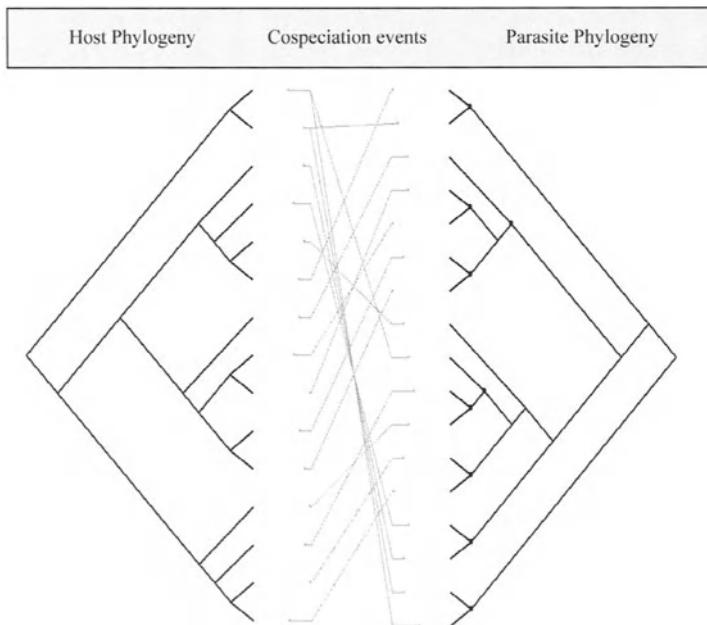
## 4.2 Correlates of cladogenesis

A complete phylogeny coupled with trait databases can be used to investigate correlates of cladogenesis or “key innovations”. Several

biological and ecological traits have been shown to cause lineages to become more diverse: phytophagy in insects (Mitter *et al.*, 1988), sexual dichromatism in birds (Barraclough *et al.*, 1995; Owens *et al.*, 1999), and polyandry in insects (Arnqvist *et al.*, 2000) are some examples (but for a counterexample of sexual selection and speciation rate in birds, see Morrow *et al.*, 2003). Until recently, rigorous hypothesis testing was limited to discrete characters compared across sister clades (reviewed in Barraclough *et al.*, 1998). However, the types of questions that can be addressed have exploded with the development of statistical methods for analyzing continuous character states simultaneously in a nested hierarchy of phylogenetically independent comparisons (Isaac *et al.*, 2003). The effect of continuous characters such as body size, sexual dimorphism, and group size and composition on species-richness now can be addressed rigorously, but only in a framework of complete phylogenetic information. Tests using supertrees are revealing some patterns for what factors underlie differences in species richness (e.g., Gittleman and Purvis, 1998; Katzourakis *et al.*, 2001; Orme *et al.*, 2002; Salamin and Davies, 2004; Isaac *et al.*, in prep.). For example, the most prevalent hypothesis is that high species richness is associated with small body size because habitats, reproduction, or dietary flexibility will promote high diversity in clades containing small-bodied species. The first phylogenetic test of this hypothesis (Gittleman and Purvis, 1998) was made possible by the existence of the primate and carnivore supertrees, and showed that there was little evidence for an association between body mass and species diversity in these groups.

## 5. Co-speciation

Interest has focused recently on investigating the patterns of joint speciation of two or more lineages; the best example is that between parasites and their hosts (reviewed in Page and Charleston, 1998; Page, 2003; Figure 3). Researchers are examining how well such lineages track each other; for example, if parasites track their host perfectly, then the respective trees would be expected to be mirror images of one another. If processes other than co-speciation occur, then parasites might switch lineages or speciate independently of the host (Page, 2003). Understanding the patterns of co-speciation can lead to understanding the process of adaptation of parasites and their hosts and the relative rate of evolution of these two clades. For example, Morand *et al.* (2000) demonstrated that louse body size was dependent on the size of their gopher hosts through a lock-and-key relationship that depended upon the thickness of the hair of the gopher and the size of the groove on the head of the louse with which it grips the hair.



**Figure 3.** Mapping co-speciation. Estimating the degree of congruence between two phylogenies is possible with complete trees of both clades (e.g., hosts and parasites).

The majority of recent studies seem to indicate a difference in congruence between interacting lineages: congruence is imperfect or absent for most kind of interactions, although there seems to be stronger support for associations between intracellular bacterial symbionts and their invertebrate hosts (e.g., Clark *et al.*, 2000). Complete trees seem to be crucial for these studies to accurately map co-speciation events; these analyses require not just one, but two complete trees. The development of new statistical methods to incorporate the fact that parasites can transfer horizontally onto hosts and that both phylogenies might contain topological errors makes this an exciting area of future research (Charleston, 1998; Huelsenbeck *et al.*, 2003).

## 6. Community ecology

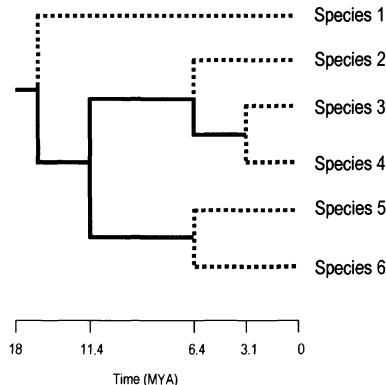
Interest has grown recently in using phylogenies to investigate ecological community structure (Tofts and Silvertown, 2000; Webb, 2000; Webb and Pitman, 2002; reviewed in Webb *et al.*, 2002). Differences in community structure will reflect not only the similarity of the environments, but also the phylogenetic similarity of the lineages within the two communities. For

example, Webb (2000) compared tree species found within 0.16 ha plots of Indonesian rain forest with species drawn randomly from the species pool. Phylogenetic relatedness of the co-occurring species was found to be higher than that of those selected from random, suggesting some degree of phylogenetic niche conservatism.

In addressing these community level questions, complete trees are less essential than having a tree that covers all the species within the community or its source pool (i.e., species present at a scale larger than the community being studied from which the community can be assembled). However, using complete trees might still be important because species within communities interact on the basis of phenotype, and incomplete taxon sampling could lead to inaccurate inferences of character evolution and assessment of which traits are conserved or homoplastic. Generating such inclusive trees that sample complete communities (“community phylogenies”; Webb *et al.*, 2002) seems viable currently only using supertree construction techniques.

## 7. Biodiversity and conservation

Species biodiversity is being lost at an alarming rate. For example, according to the latest figures assembled by the IUCN (2002 IUCN Red List of Threatened Species; <http://www.redlist.org>), at least one-third of mammals are threatened with extinction. Applications of complete trees have an important role to play in conservation because they can give some answers to questions that otherwise would not be accessible from traditional phylogenies (Mace *et al.*, 2003). Complete trees that have estimates of taxon divergence times can be used to estimate measures of phylogenetic diversity (*PD*) from the lengths of the branches separating different taxa (Faith, 1992; Nee and May, 1997). Dated supertrees allow a measure of *PD* for a large clade of hundreds of species that would not be possible from other, conventional phylogenies. For example, in the Chiroptera, the bumblebee bat (*Craseonycteris thonglongyai*) is the sole representative of a distinctive family of bats that originated approximately 40 million years ago (Jones *et al.*, in prep.). This species will have a higher measure of phylogenetic diversity, and therefore might be more worthy of conservation effort, than a species of *Myotis* bat which comes from a genus of over 60 species that originated approximately 10 million years ago (Figure 4). These estimates of *PD* can be used in different ways to inform us of the potential impact of the current extinction crisis and to help inform policy makers of the best ways to ameliorate human impacts on biodiversity.



**Figure 4.** Losing shared and unique evolutionary history. Phylogenetic diversity (PD) is shown by the length of the branches (in units of time) joining different species in the phylogeny. PD that is shared between all species in the clade is shown in solid lines, whereas PD that is unique to each species is shown in dashed lines. The extinction of species 1 would have a greater impact on the phylogenetic diversity of the clade than if species 3 was lost (adapted from Mace *et al.*, 2003).

## 7.1 Impacts of current extinction crisis

Interest in the potential effects of extinction on the Tree of Life were sparked by a simulation study showing that the majority of phylogenetic diversity in the Tree of Life would be preserved (81%) under specific conditions (e.g., cladogenesis in the form of asymmetric trees and random extinction), even with high numbers of species extinctions (95%) (Nee and May, 1997). However, extinction is not random: closely related taxa share similar levels of threat (Russell *et al.*, 1998; see also Mace *et al.*, 2003), such that when one species is threatened it is probable that closely related taxa in the tree are also threatened. Complete supertrees for primates and carnivores were used to estimate the amount of PD lost if species classified as threatened according to the 2002 IUCN Red List of Threatened Species (<http://www.redlist.org>) went extinct relative to random extinction (Purvis *et al.*, 2000a; see also Mooers *et al.*, in press). In primates, but not in carnivores, significantly more PD would be lost than expected. Primates have a tree that is more unbalanced than carnivores (many long branches with few species and short branches with many species), thus making the effect of the non-random distribution of threat more pronounced. This points to another crucial reason for using a complete tree: tree balance would not have been possible to estimate, nor would its effects on these results.

## 7.2 Conservation priority setting

Conservation policy makers are interested in setting priorities in the face of limited resources to minimize the impacts of the human-caused extinction crisis. Many conservation priority-setting exercises are area or species based, focusing on distinctive areas or species to preserve as much biological diversity as possible (i.e., conservation hotspots; Myers *et al.*, 2000). *PD* is another measure of distinctiveness that is starting to be recognized as being important to conservation-policy decisions (Purvis *et al.*, in press). For example, it is important to know how if areas that are important for total numbers of species (species-richness hotspots) are those that also contain the most phylogenetic diversity (*PD* hotspots). Evidence would suggest that these hotspots are essentially overlapping (Sechrest *et al.*, 2002), but more complete dated trees for different clades are needed to address these questions comprehensively. Such trees can be used additionally to indicate the processes that have created the pattern of current biodiversity. For example, the phylogeny could be used to identify rapidly diversifying clades, which combined with geographic information, might enable us to distinguish between “cradles” of diversity from “museums” (Chown and Gaston, 2000; Mace *et al.*, 2003).

## 8. Caveats to supertree usage

This chapter rests on a key assumption: supertrees are an effective means for procuring complete, large trees that are very difficult, if not impossible, to acquire from traditional methods. Therefore, it is important to assess whether the shortcut of a supertree is indeed effective in the sense of providing phylogenetic accuracy. As described above, there are many exciting reasons to use supertree construction techniques for comparative hypothesis testing. Nevertheless, similar to using phylogenies in general, supertrees can give inaccurate phylogenetic information for comparative studies. The following problems are apparent with the supertrees that have been used thus far; solutions for dealing with these problems are preliminary because we have started using supertrees only recently.

The first issue is whether branch lengths are wrong in dated supertrees (see also Moore *et al.*, 2004). Errors could involve using inaccurate molecular information or when there is clear disparity between molecular divergence times and fossil evidence (assuming the rarity of a molecular phylogeny including fossil taxa). In either case, all comparative tests will suffer from these problems. Another form of error could arise from missing taxa. The influence missing taxa have on a comparative result depends on

where omissions occur in the phylogeny relative to the clade that is being tested. Some diagnostic tests can be performed to investigate lineages-through-time plots to see if there are numerous missing taxa at certain time intervals. If the pattern does not match up with the known extinction patterns for a lineage, then there are probably missing taxa that should be accounted for or tested against alternative trees. Similar to phylogenetic tests in general (Losos, 1994; Martins, 1996; Housworth and Martins, 2001), an analytical solution is to generate alternative phylogenies in which the branch lengths are varied to examine the statistical power of a comparative result.

Perhaps the most worrisome issue results from the “garbage in, garbage out” phenomenon, where conflicting or wrong information is used for supertree construction. There are many statistical methods for diagnosing this problem (Bininda-Emonds *et al.*, 2002), but all involve essentially evaluating the kind of information contributing to uncertainty in the tree topology. The solution, again, is to execute the comparative test against alternative tree topologies to examine what the probability is of incorrect information giving a result. There is another solution, however. When it is known that portions of the supertree are reliable but others are inaccurate, as is frequently the case in supertrees where the relationships of taxa are poorly represented or only in the form of taxonomic rank (see Bininda-Emonds *et al.*, 1999; Salamin *et al.*, 2002), one method is to constrain certain nodes that are reliable and then reorder the remaining nodes through a randomization process (Housworth and Martins, 2001). This process would acknowledge in that some of the tree is useful a comparative analysis, but then assess statistically how different topologies influence a result. Although simulation studies show clearly that using even only a small amount of phylogenetic information is better than none (Gittleman and Luh, 1992; Losos, 1994), the reordering method does not handle error in branch lengths or an expected model of character evolution along the branches.

For comparative tests of particular traits, supertrees might bias results because of the different kinds of information they rest upon. For example, a comparative analysis of brain size evolution testing for differences in rate change across clades could be influenced by a supertree that uses disproportionately more phylogenies based on skull characters than molecular ones. A useful diagnostic would be to assess whether the phylogenies that are influential in a supertree are based on the same traits that are the focus of comparative study.

Finally, inaccuracy within a supertree might not always be detrimental. In the context of using supertrees to catalogue what is known about a taxon phylogenetically, many clades that are known poorly are precisely those that are known poorly in general and the most threatened by extinction (see Mace *et al.*, 2003). For example, across the carnivore supertree, Bremer support

values are the lowest for the Herpestidae, a family of 37 mongoose species, of which little information is available on even body size or geographic range distribution. Supertrees, therefore, can be used to send a forceful and empirical signal about which clades desperately need more study.

## 9. Future directions

What is the fate of supertrees? Undoubtedly the answer will change as primary sources for phylogenies become assembled more systematically, cover a greater span of characters across clades, and are comprehensive across large taxonomic groups. Admittedly, the future of supertrees might be limited, if not obsolete, when complete phylogenies are achieved by other consensus methods. Yet, we see two areas that might be long lasting for supertrees, even when other sources of phylogenetic material become available. One is the inevitable problem of achieving completeness in studies of planetary biodiversity. About 1.5 to 1.8 million species have been described, with anywhere from 3.6 to 100 million remaining to be discovered (Wilson, 2002). Only a fraction of these are known phylogenetically, even for groups such as the mammals that are very well studied (see Bininda-Emonds *et al.*, 1999; Jones *et al.*, 2002). As species concepts become more blurred (see Agapow *et al.*, 2004), it will be more important to develop a measure of biodiversity that is objective and consistent across diverse clades (Mace *et al.*, 2003). A phylogenetically-based concept will be helpful (Purvis and Hector, 2000; Mace *et al.*, 2003), but only if we have phylogenetic information. As shown above for studies of biodiversity and conservation, supertrees will give us phylogenetic information that is available and, more importantly, tell us which clades are crucial to go and gather this information.

The harshest criticism of supertrees comes from the opinion that consistent, full phylogenies must rely on data compilation methods of characters rather than phylogenies (e.g., Gatesy *et al.*, 2002; Gatesy and Springer, 2004). But what if no single character type (i.e., given gene or region of body) will ever be measured for an entire large clade? This is the case for one ultra-charismatic group: the dinosaurs. Phylogenetic information for dinosaurs will never be complete because they are known unfortunately only from a very fragmented fossil record. Therefore, all phylogenies must be derived from character pieces — the essence of what supertrees are based on from individual phylogenies. If nothing else, a complete supertree of all dinosaurs (see Pisani *et al.*, 2002) will be a legacy contribution of supertrees if it provides the best possible phylogeny of this charismatic group.

## Acknowledgements

We thank Olaf Bininda-Emonds for the invitation to contribute a chapter to this book; the Editor, Arne Mooers and an anonymous reviewer for helpful suggestions on improving the manuscript; and the National Science Foundation for support (grant DEB/0129009). This work was completed while KEJ was at the Department of Biology, University of Virginia.

## References

- ABOUHEIF, E. 1999. A method for testing the assumption of phylogenetic independence in comparative data. *Evolutionary Ecology Research* 1:895–909.
- AGAPOV, P.-M., BININDA-EMONDS, O. R. P., CRANDALL, K. A., GITTLEMAN, J. L., MACE, G. M., MARSHALL, J. C., AND PURVIS, A. In press. The impact of species concept on biodiversity studies. *Quarterly Review of Biology*.
- ALLEN, A. P., BROWN, J. H., AND GILLOOLY, J. F. 2002. Global biodiversity, biochemical kinetics and the energetic-equivalence rule. *Science* 297:1545–1548.
- ARNQVIST, G., EDVARDSSON, M., FRIBERG, U., AND NILSSON, T. 2000. Sexual conflict promotes speciation in insects. *Proceedings of the National Academy of Sciences of the United States of America* 97:10460–10464.
- BARRACLOUGH, T. G., HARVEY, P. H., AND NEE, S. 1995. Sexual selection and taxonomic diversity in passerine birds. *Proceedings of the Royal Society London B* 259:211–215.
- BARRACLOUGH, T. G. AND NEE, S. 2001. Phylogenetics and speciation. *Trends in Ecology and Evolution* 16:391–399.
- BARRACLOUGH, T. G. AND VOGLER, A. P. 2000. Detecting the geographical pattern of speciation from species-level phylogenies. *American Naturalist* 155:419–434.
- BARRACLOUGH, T. G., VOGLER, A. P., AND HARVEY, P. H. 1998. Revealing the factors that promote speciation. *Philosophical Transactions of the Royal Society* 353:241–249.
- BARTON, R. A. 1998. Visual specialization and brain evolution in primates. *Proceedings of the Royal Society of London B* 265:1933–1937.
- BARTON, R. A. AND HARVEY, P. H. 2000. Mosaic evolution of brain structure in mammals. *Nature* 405:1055–1058.
- BININDA-EMONDS, O. R. P. In press. The phylogenetic position of the giant panda (*Ailuropoda melanoleuca*): a historical consensus through supertree analysis. In D. G. Lindburg and K. Baragona (eds), *Pandas: Biology and Conservation*. University of California Press, Berkeley.
- BININDA-EMONDS, O. R. P. AND GITTLEMAN, J. L. 2000. Are pinnipeds functionally different from fissiped carnivores? The importance of phylogenetic comparative analyses. *Evolution* 54:1011–1023.
- BININDA-EMONDS, O. R. P., GITTLEMAN, J. L., AND PURVIS, A. 1999. Building large trees by combining phylogenetic information: a complete phylogeny of the extant Carnivora (Mammalia). *Biological Reviews* 74:143–175.
- BININDA-EMONDS, O. R. P., GITTLEMAN, J. L., AND STEEL, M. A. 2002. The (super)tree of life: procedures, problems, and prospects. *Annual Review of Ecology and Evolution* 33:265–289.
- BLOMBERG, S. P., GARLAND, T., AND IVES, A. R. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution* 57:717–745.

- BRYANT, D., SEMPLE, C., AND STEEL, M. 2004. Supertree methods for ancestral divergence dates and other applications. In O. R. P. Bininda-Emonds (ed). *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 129–150. Kluwer Academic, Dordrecht, the Netherlands.
- CHARLESTON, M. A. 1998. Jungles: a new solution to the host/parasite phylogeny reconciliation problem. *Mathematical Biosciences* 149:191–223.
- CHEVERUD, J. M., DOW, M. M., AND LEUTENEGGER, W. 1985. The quantitative assessment of phylogenetic constraints in comparative analyses: sexual dimorphism in body weight among primates. *Evolution* 39:1335–1351.
- CHOWN, S. L. AND GASTON, K. J. 2000. Areas, cradles and museums: the latitudinal gradient in species richness. *Trends in Ecology and Evolution* 15:311–315.
- CLARK, M. A., MORAN, N. A., BAUMAN, P., AND WERNEGREN, J. J. 2000. Cospeciation between bacterial endosymbionts (*Buchnera*) and a recent radiation of aphids (*Uroleucon*) and pitfalls of testing for phylogenetic congruence. *Evolution* 54:517–525.
- DARWIN, C. 1859. *The Origin of Species*. London, John Murray.
- DEANER, R. O. AND NUNN, C. L. 1999. How quickly do brains catch up with bodies? A comparative method for detecting evolutionary lag. *Proceedings of the Royal Society of London Series B* 266:687–694.
- FELSENSTEIN, J. 1985. Phylogenies and the comparative method. *American Naturalist* 125:1–15.
- FAITH, D. P. 1992. Conservation evaluation and phylogenetic diversity. *Biological Conservation* 61:1–10.
- FRECKLETON, R. P., HARVEY, P. H., AND PAGEL, M. D. 2002. Phylogenetic analysis and comparative data: a test and review of the evidence. *American Naturalist* 160:712–726.
- GASTON, K. J. 1998. Species range size distributions: products of speciation, extinction and transformation. *Philosophical Transactions of the Royal Society* 353:219–230.
- GATESY, J., MATTHEE, C., DESALLE, R., AND HAYASHI, C. 2002. Resolution of a supertree / supermatrix paradox. *Systematic Biology* 51:652–664.
- GATESY, J. AND SPRINGER, M. S. 2004. A critique of matrix representation with parsimony supertrees. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 369–388. Kluwer Academic, Dordrecht, the Netherlands.
- GITTLEMAN, J. L. AND KOT, M. 1990. Adaptation: statistics and a null model for estimating phylogenetic effects. *Systematic Zoology* 39:227–241.
- GITTLEMAN, J. L. AND LUH, H.-K. 1992. On comparing comparative methods. *Annual Review of Ecology and Systematics* 23:383–404.
- GITTLEMAN, J. L. AND PURVIS, A. 1998. Body size and species-richness in carnivores and primates. *Proceedings of the Royal Society of London B* 265:113–119.
- HARVEY, P. H. AND PAGEL, M. D. 1991. *The Comparative Method in Evolutionary Biology*. Oxford University Press, Oxford.
- HARVEY, P. H. AND RAMBAUT, A. 2000. Comparative analyses for adaptive radiations. *Philosophical Transactions of the Royal Society* 355:1–7.
- HILLIS, D. M. 1996. Inferring complex phylogenies. *Nature* 383:130–131.
- HOUSWORTH, E. A. AND MARTINS, E. P. 2001. Random sampling of constrained phylogenies: conducting phylogenetic analyses when the phylogeny is partially known. *Systematic Biology* 50:628–639.
- HUELSENBECK, J. P., RANNALA, B., AND LARGET, B. 2003. A statistical perspective for reconstructing the history of host-parasite associations. In R. D. M. Page (ed.), *Tangled*

- Trees. Phylogeny, Cospeciation and Coevolution*, pp. 93–119. Chicago University Press, Chicago and London.
- ISAAC, N. J. B., AGAPOW, P.-M., HARVEY, P. H., AND PURVIS, A. 2003. Phylogenetically nested comparisons for testing continuous correlates of species richness: a simulation study. *Evolution* 57:18–26.
- JOHNSON, C. N. 1998. Species extinction and the relationship between distribution and abundance. *Nature* 394:272–274.
- JONES, K. E., PURVIS, A., AND GITTLEMAN, J. L. 2003. Biological correlates of extinction risk in bats. *American Naturalist* 161:601–614.
- JONES, K. E., PURVIS, A., MACLARNON, A., BININDA-EMONDS, O. R. P., AND SIMMONS, N. B. 2002. A phylogenetic supertree of the bats (Mammalia: Chiroptera). *Biological Reviews* 77:223–259.
- JONES, K. E., SECHREST, W., AND GITTLEMAN, J. L. In press. Geography and phylogeny: identifying global patterns and implications for conservation. In A. Purvis, J. L. Gittleman, and T. M. Brooks (eds), *Phylogeny and Conservation*. Cambridge University Press, Cambridge.
- KATZOURAKIS, A., PURVIS, A., AZMEH, S., ROTHEROW, G., AND GILBERT, F. 2001. Macroevolution of hoverflies (Diptera: Syrphidae): the effect on the use of higher level taxa in studies of biodiversity and correlates of species richness. *Journal of Evolutionary Biology* 14:219–227.
- KIRKPATRICK, M. AND SLATKIN, M. 1993. Searching for evolutionary patterns in the shape of a phylogenetic tree. *Evolution* 47:1171–1181.
- LANDE, R. 1979. Quantitative genetic analysis of multivariate evolution, applied to brain:body size allometry. *Evolution* 33:402–416.
- LINDER, H. P. 2000. Vicariance, climate change, anatomy and phylogeny of Restionaceae. *Biological Journal of the Linnean Society* 134:159–177.
- LIU, F.-G. R., MIYAMOTO, M. M., FREIRE, N. P., ONG, P. Q., TENNANT, M. R., YOUNG, T. S., AND GUGEL, K. F. 2001. Molecular and morphological supertrees for eutherian (placental) mammals. *Science* 291:1786–1789.
- LOSOS, J. B. 1994. An approach to the analysis of comparative data when a phylogeny is unavailable or incomplete. *Systematic Biology* 43:117–123.
- LOSOS, J. B. 2000. Ecological character displacement and the study of adaptation. *Proceedings of the National Academy of Sciences of the United States of America* 97:5693–5695.
- LOSOS, J. B. AND GLOR, R. E. 2003. Phylogenetic comparative methods and the geography of speciation. *Trends in Ecology and Evolution* 18:220–227.
- MACE, G. M., GITTLEMAN, J. L., AND PURVIS, A. 2003. Preserving the Tree of Life. *Science* 300:1707–1709.
- MARTINS, E. P. 1996. Conducting phylogenetic comparative studies when the phylogeny is not known. *Evolution* 50:12–22.
- MARTINS, E. P., DINIZ-FILHO, J. A. F., AND HOUSWORTH, E. A. 2002. Adaptive constraints and the phylogenetic comparative method: a computer simulation test. *Evolution* 56:1–13.
- MITTER, C., FARRELL, B., AND WIEGMANN, B. 1988. The phylogenetic study of adaptive zones: has phytophagy promoted insect diversification? *American Naturalist* 132:107–128.
- MOOERS, A. Ø., HEARD, S. B., AND CHROSTOWSKI, E. In press. Evolutionary heritage as a metric for conservation. In A. Purvis, J. L. Gittleman, and T. M. Brooks (eds), *Phylogeny and Conservation*. Cambridge University Press, Cambridge.
- MOORE, B. R., CHAN, K. M. A., AND DONOGHUE, M. J. 2004. Detecting diversification rate variation in supertrees. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees*:

- Combining Information to Reveal the Tree of Life*, pp. 487–533. Kluwer Academic, Dordrecht, the Netherlands.
- MORAND, S., HAFNER, M. S., PAGE, R. D. M., AND REED, D. L. 2000. Comparative body size relationships in pocket gophers and their chewing lice. *Biological Journal of the Linnean Society* 70:239–249.
- MORROW, E. H., PITCHER, T. E., AND AMQVIST, G. 2003. No evidence that sexual selection is an ‘engine of speciation’ in birds. *Ecology Letters* 6:228–234.
- MYERS, N., MITTERMEIER, R. A., MITTERMEIER, C. G., DA FONSECA, G. A. B., AND KENT, J. 2000. Biodiversity hotspots for conservation priorities. *Nature* 403:853–858.
- NEE, S. 2001. Inferring speciation rates from phylogenies. *Evolution* 55:661–668.
- NEE, S. AND MAY, R. M. 1997. Extinction and the loss of evolutionary history. *Science* 278:692–694.
- NEE, S., MAY, R. M., AND HARVEY, P. H. 1994. The reconstructed evolutionary process. *Philosophical Transactions of the Royal Society London B* 344:305–311.
- NEE, S., MOOERS, A. Ø., AND HARVEY, P. H. 1992. Tempo and mode of evolution revealed from molecular phylogenies. *Proceedings of the National Academy of Sciences of the United States of America* 89:8322–8326.
- NUNN, C. L., GITTLEMAN, J. L., AND ANTONOVICS, J. 2000. Promiscuity and the primate immune system. *Science* 290:1168–1170.
- NUNN, C. L., GITTLEMAN, J. L., AND ANTONOVICS, J. 2003. A comparative study of white blood cell counts and disease risk in carnivores. *Proceedings of the Royal Society London B* 270:347–356.
- ORME, C. D. L., ISAAC, N. J. B., AND PURVIS, A. 2002. Are most species small? Not within species-level phylogenies. *Proceedings of the Royal Society London B* 269:1279–1287.
- OWENS, I. P. F., BENNETT, P. M., AND HARVEY, P. H. 1999. Species richness among birds: body size, life history, sexual selection or ecology? *Proceedings of the Royal Society of London B* 266:933–939.
- PAGE, R. D. M. 2003. *Tangled Trees: Phylogeny, Cospeciation and Coevolution*. University of Chicago Press, Chicago and London.
- PAGE, R. D. M. AND CHARLESTON, M. A. 1998. Trees within trees: phylogeny and historical associations. *Trends in Ecology and Evolution* 13:356–359.
- PAGEL, M. D. 1999. Inferring the historical patterns of biological evolution. *Nature* 401:877–884.
- PAGEL, M. D. AND HARVEY, P. H. 1988. How mammals produce large-brained offspring. *Evolution* 42:948–957.
- PARADIS, E. 1998. Detecting diversification rates without fossils. *American Naturalist* 152:176–187.
- PISANI, D., YATES, A. M., LANGER, M. C., AND BENTON, M. J. 2002. A genus-level supertree of the Dinosauria. *Proceedings of the Royal Society of London B Biological Sciences* 269:915–921.
- PRICE, T. 1997. Correlated evolution and independent contrasts. *Philosophical Transactions of the Royal Society London Series B* 352:519–529.
- PURVIS, A. 1995. A composite estimate of primate phylogeny. *Philosophical Transactions of the Royal Society London Series B* 348:405–421.
- PURVIS, A., AGAPOW, P.-M., GITTLEMAN, J. L., AND MACE, G. M. 2000a. Non-random extinction increases the loss of evolutionary history. *Science* 288:328–330.
- PURVIS, A., GITTLEMAN, J. L., AND BROOKS, T. M. (eds). In press. *Phylogeny and Conservation*. Cambridge University Press, Cambridge.

- PURVIS, A., GITTLEMAN, J. L., COWLISHAW, G., AND MACE, G. M. 2000b. Predicting extinction risk in declining species. *Proceedings of the Royal Society of London B* 267:1947–1952.
- PURVIS, A., GITTLEMAN, J. L., AND LUH, H. K. 1994. Truth or consequences: effects of phylogenetic accuracy on two comparative methods. *Journal of Theoretical Biology* 167:293–300.
- PURVIS, A., NEE, S., AND HARVEY, P. H. 1995. Macroevolutionary inferences from primate phylogeny. *Proceedings of the Royal Society of Science London B* 260:329–333.
- PURVIS, A. AND HECTOR, A. 2000. Getting the measure of biodiversity. *Nature* 405:212–219.
- PURVIS, A., WEBSTER, A. J., AGAPOW, P.-M., JONES, K. E., AND ISAAC, N. J. B. 2003. Primate life histories and phylogeny. In P. M. Kappeler and M. E. Pereira (eds) *Primate Life Histories and Socioecology*, pp. 25–40. Chicago University Press, Chicago and London.
- PYBUS, O. G. AND HARVEY, P. H. 2000. Testing macroevolutionary models using incomplete molecular phylogenies. *Proceedings of the Royal Society of Science London B* 267:2267–2272.
- PYBUS, O. G., RAMBAUT, A., HOLMES, E. C., AND HARVEY, P. H. 2002. New inferences from tree shape: numbers of missing taxa and population growth rates. *Systematic Biology* 51:881–888.
- RUSSELL, G. J., BROOKS, T. M., MCKINNEY, M. M., AND ANDERSON, C. G. 1998. Present and future taxonomic selectivity in bird and mammal extinctions. *Conservation Biology* 12:1365–1376.
- SALAMIN, N. AND DAVIES, T. J. 2004. Using supertrees to investigate species richness in grasses and flowering plants. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 461–486. Kluwer Academic, Dordrecht, the Netherlands.
- SALAMIN, N., HODKINSON, T. R., AND SAVOLAINEN, V. 2002. Building supertrees: an empirical assessment using the grass family (Poaceae). *Systematic Biology* 51:136–150.
- SANDERSON, M. J. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* 19:301–302.
- SECHREST, W., BROOKS, T. M., DA FONSECA, G. A. B., KONSTANT, W. R., MITTERMEIER, R. A., PURVIS, A., RYLANDS, A. B., AND GITTLEMAN, J. L. 2002. Hotspots and the conservation of evolutionary history. *Proceedings of the National Academy of Sciences of the United States of America* 99:2067–2071.
- SLOWINSKI, J. B. AND GUYER, C. 1989. Testing null models in questions of evolutionary success. *Systematic Zoology* 38:189–191.
- TOFTS, R. AND SILVERTOWN, J. 2000. A phylogenetic approach to community assembly from a local species pool. *Proceedings of the Royal Society of Science London B* 267:363–369.
- VOS, R. A. AND MOOERS, A. Ø. 2004. Reconstructing divergence times for supertrees. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 281–299. Kluwer Academic, Dordrecht, the Netherlands.
- WEBB, C. O. 2000. Exploring the phylogenetic structure of ecological communities: an example for rain forest trees. *American Naturalist* 156:145–155.
- WEBB, C. O. AND PITMAN, N. C. A. 2002. Phylogenetic balance and ecological evenness. *Systematic Biology* 51:898–907.
- WEBB, C. O., ACKERLY, D. D., MCPEEK, M. A., AND DONOGHUE, M. J. 2002. Phylogenies and community ecology. *Annual Review of Ecology and Systematics* 33:475–505.
- WEBB, T. J. AND GASTON, K. J. 2000. Geographic range size and evolutionary age in birds. *Proceedings of the Royal Society London B* 267:1843–1850.
- WILSON, E. O. 2002. *The Future of Life*. Knopf, New York.

## Chapter 21

# USING SUPERTREES TO INVESTIGATE SPECIES RICHNESS IN GRASSES AND FLOWERING PLANTS

Nicolas Salamin and T. Jonathan Davies

**Abstract:** Matrix representation with parsimony is the most widely used method for supertree reconstruction, due mainly to its ability to deal with incompatible source trees, and its simple and logical mathematical basis. Supertrees have the advantage over consensus methods in that the source trees do not need to contain identical terminal taxa, but only overlap. This makes supertrees a useful and attractive approach to building comprehensive phylogenetic trees, which are indispensable tools for investigating macroevolutionary patterns. Here, we highlight the use of supertrees of two plant lineages. We used the genus-level supertree of grasses (containing almost two-thirds of grass genera) and a family-level supertree of the angiosperms to investigate the influence of various putative key innovations (habit, life form, sex, mode of pollination, mode of dispersal, water resistance, salt tolerance, and habitat preference) on species richness at two different taxonomic levels within the flowering plants. The results suggest that no significant increase in speciation rates could be linked to any of these features in the angiosperms, whereas life form had a significant impact on the number of species at the family level in the grasses.

**Keywords:** angiosperms; grasses; key innovations; macroevolution; speciation

### 1. Introduction

Knowledge of the evolutionary history among groups of taxa is an essential element for classification purposes and investigations of macroevolutionary processes. Areas such as genomics (Dacks and Doolittle, 2001; Koch *et al.*, 2001; Soltis *et al.*, 2002), developmental biology (Halanych and

*Bininda-Emonds, O. R. P. (ed.) Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life, pp. 461–486. Computational Biology, volume 3 (Dress, A., series ed.). © 2004 Kluwer Academic Publishers.*

Passamaneck, 2001; Jeffery *et al.*, 2002; Simpson, 2002), and ecology (Christensen *et al.*, 2002; Foley, 2002; Schmid-Hempel and Ebert, 2003) are also now taking advantage of the information contained in phylogenetic trees.

It has become apparent that building comprehensive phylogenetic trees is of paramount importance in evolutionary studies. Accurate inferences of macroevolutionary processes require most of the diversity of taxa within a group be sampled, either to reduce the chance of misleading the tree-reconstruction process (Graybeal, 1998; Hillis, 1998; Zwickl and Hillis, 2002) or to encapsulate most of the relevant information to make optimal use of the evolutionary history obtained (Purvis, 1996). For large groups of organisms containing tens or hundreds of thousands of species (such as the flowering plants or the grasses), the task of sampling an adequate number of taxa and gathering sufficient information to build a phylogenetic tree can easily become an immense and costly task.

The revolution in molecular techniques has eased the production of DNA sequences, and studies containing more than a hundred taxa based on molecular characters are now commonplace (e.g., Chase *et al.*, 1993; Källersjö *et al.*, 1998; Soltis *et al.*, 1999; Qiu *et al.*, 2000; Savolainen *et al.*, 2000; Marvaldi *et al.*, 2002). However, methods based on primary biological characters are not always applicable on a larger scale, principally because of uncoordinated data collection resulting in either a patchwork of coverage for a given taxonomic group or the extensive use of only a few types of DNA sequences (Bininda-Emonds *et al.*, 2002). In such cases, a meta-analysis approach as used in supertree reconstruction might be more appropriate.

Supertree reconstruction takes advantage of existing, less comprehensive phylogenetic trees, and assembles them into a coherent and accurate representation of the relationships among the whole set of taxa at hand. Unlike consensus methods, it can deal with trees that do not have the same set of terminal taxa. It is, therefore, able to build a more comprehensive phylogenetic tree than any represented in the individual source trees on which it is based. Different methods, classified as being either “direct” or “indirect” (Wilkinson *et al.*, 2001; see also Wilkinson *et al.*, 2004), exist to build supertrees. Here, we concentrate on the “indirect” matrix representation with parsimony (MRP; Baum, 1992; Ragan, 1992) method and we refer the reader to the given references and Baum and Ragan (2004) for more technical information on the method. MRP is a method that is able to deal with source trees containing incompatible nodes without necessitating a loss of resolution. Its simple logical and mathematical basis, coupled with its ease of implementation makes it the most commonly used method in supertree reconstruction at the moment (Bininda-Emonds *et al.*, 2002). As a consequence of their potential for complete taxonomic coverage,

supertrees have been applied to a broad range of ecological and evolutionary analyses (see Bininda-Emonds *et al.*, 2002; Gittleman *et al.*, 2004). For example, they have been used to study rates of cladogenesis (Purvis *et al.*, 1995; Paradis, 1998; Bininda-Emonds *et al.*, 1999), functional relationships (e.g., Johnson, 1998; Linder, 2000; Nunn and Barton, 2000), and to associate differences in species richness with changes in phenotypic traits (Gittleman and Purvis, 1998). Their inclusive taxonomic coverage increases the power of such tests and lessens the effect of incomplete taxon sampling on comparative analyses (Bininda-Emonds *et al.*, 2002).

In investigations into patterns of species richness, it has been proposed that certain traits might influence the rate of evolution and the production of new species. Such traits have been called key innovations, and it has been suggested that these traits have enabled those lineages that possess them to proliferate at an increased rate by opening up new adaptive zones (Burger, 1981; Maynard Smith and Szathmary, 1995). If this model is correct, then the expectation is that the observed differences in species richness between certain clades are correlated with the presence of particular traits. In attempting to identify correlates of species richness, the hierarchical nature of evolutionary history means that treating taxa as independent evolutionary units can result in erroneous inferences (Felsenstein, 1985). Therefore, a phylogenetic approach is essential in such studies.

Here, we used supertree reconstructions for both the grass family and families of angiosperms to investigate the effect of diverse phenotypic traits on the species richness within these two plant lineages. Factors that potentially influence species richness have been the focus of intensive study. Our goal in this study was not to present new methodology or results, but rather to extend the approach to a much larger sample of taxa. Supertrees are ideally suited in our case because they both provide the capability of combining different sets of taxa into a more comprehensive analysis and put these taxa into a phylogenetic context, which is extremely important in macroevolutionary studies. We first investigate factors that could have promoted speciation in the grasses using the largest phylogeny for the grass family, a genus-level supertree containing two-third of the grass genera (Salamini *et al.*, 2002). Five traits reviewed in the following section were analyzed as well as traits that are thought to have been important in grass evolution (e.g., drought resistance, salt tolerance, and adaptation to open habitat). Additionally, a supertree for the angiosperms containing the most complete set of angiosperm families (Davies *et al.*, in press) was used to investigate the pattern of species richness in the flowering plants as a whole and the effect of the five putative key innovations described in the following section on the rates of diversification.

## 2. Phenotypic traits and species richness

Different aspects of plant biology have been suggested to influence the rate of speciation in angiosperms. Numerous studies have investigated putative key characters thought to be correlated with increased species richness in angiosperms with mixed success and with varying degrees of phylogenetic rigor (e.g., Burger, 1981; Farrell *et al.*, 1991; Marzluff and Dial, 1991; Eriksson and Bremer, 1992; Manning and Linder, 1992; Ricklefs and Renner, 1994; Sanderson and Donoghue, 1994; Gaut *et al.*, 1996; Dodd *et al.*, 1999; Heilbutch, 2000; Silvertown *et al.*, 2000). To date, there has been little consensus on the relative importance of the five main traits examined in this study in explaining differences in the rate of speciation between lineages. We now review briefly some of the principle hypotheses concerning these five traits.

### 2.1 Habit

It has been proposed that increased reproductive rate increases speciation and decreases the chance of extinction (Marzluff and Dial, 1991). Woody plants take longer to mature typically, and are thought to have longer generation times, which has been suggested to be correlated negatively to the rate of evolution (see Eriksson and Bremer, 1992; Gaut *et al.*, 1992, 1996). Barraclough and Savolainen (2001) found increased rates of molecular evolution within angiosperm families was correlated positively with species richness, suggesting that there could be a line of causality from decreased generation time to increased species richness via the effect of the former on the rate of molecular evolution. However, Rosenheim and Tabashnik (1991) argued that the exact relationship between generation time and evolutionary rate was more complicated. Furthermore, little is actually known about the number of cell replications before reproduction or about ancestral generation times within plants, nor what effect the longevity of the seed bank could have. Moreover, it is possible that germ-line mutations can occur throughout the lifetime of a plant (Bousquet *et al.*, 1992).

### 2.2 Life form

Bousquet *et al.* (1992) suggested that annuals might be able to fill new niches better as a result of more functionally important mutations being driven to fixation because of smaller population sizes. Evidence in support of this hypothesis comes from the faster rates of evolution, particularly in the accumulation of nonsynonymous mutations, observed in *rbcL* within annuals (Bousquet *et al.*, 1992). Perennials also have longer generation times so the

arguments surrounding the importance of habit will also apply. However, the finding that nonsynonymous and synonymous mutation rates in annuals varied to different degrees when compared to perennials (Bousquet *et al.*, 1992) suggests that there is a more subtle relationship between life form and rate of evolution than the simple division into annuals and perennials.

### 2.3 Sex

Based upon the assumption that monoecy is correlated with self-compatibility, it has been hypothesized that monoecious species might be more likely to speciate for several reasons. First, if hybridization occurs, species that can reproduce asexually are more likely to form a new species (Rieseberg, 1997). Second, according to Baker's law (Baker, 1955), self-compatible species have increased probability of establishment after long-range dispersal, and therefore increased speciation rates. Heilbuth (2000) found dioecy to be correlated with lower species richness, but no evidence in support of Baker's Law. Finally, dioecious species are more likely to have generalist pollinators (Bawa and Opler, 1975; Bawa, 1994), thereby inhibiting the reproductive isolation necessary for speciation.

### 2.4 Mode of pollination

Burger (1981) suggested that biotic pollination was important in the early diversification of the angiosperms by allowing outcrossing sexual reproduction in highly diverse populations of few individuals. However, it appears unlikely that this would have much impact upon established populations. Furthermore, there is evidence of insect pollination before the appearance of the angiosperms, and many diverse families within the angiosperms are predominantly wind-pollinated (e.g., Poaceae, Cyperaceae, Juncaceae, and Fagaceae; Midgley and Bond, 1991). It has also been suggested that biotic pollination is a major isolating mechanism between plant species. The occurrence of faithful pollinators could therefore increase the rate of diversification (Dodd *et al.*, 1999; Ricklefs and Renner, 1994). Pollinator-mediated reproductive isolation was demonstrated in the genus *Disperis* (Manning and Linder, 1992), where, by depositing pollen on different parts of the anatomy of a pollinator, the species became effectively reproductively isolated despite sharing a common pollinator. Gorelick (2001) suggested that biotic pollination did not in fact increase speciation rates, but influenced contemporary patterns of diversity instead by decreasing the probability of extinction, thereby resulting in the increased net speciation rates observed. He argued that biotically pollinated species were found at lower densities than abiotically pollinated species, which

reduced the chances of any single event affecting all individuals in a population. Such robustness to extinction enabled populations to endure over longer evolutionary time, increasing the likelihood of speciation. Some empirical evidence supports this proposed correlation of biotic pollination (and life form) with increased species diversity (e.g., Eriksson and Bremer, 1992).

## 2.5 Mode of dispersal (as indicated by fruit type)

It has been argued that dispersal by animals increases long-distance dispersal, thereby promoting establishment of isolated populations (Eriksson and Bremer, 1992). Such isolated populations might be more likely to diverge through genetic drift and a founder effect. However, conversely, increased long-distance dispersal could also encourage backcrossing, breaking down reproductive isolation and decreasing rates of speciation (Ricklefs and Renner, 1994). It might be that a limited migration capacity is most likely to lead to increased speciation, enabling relatively infrequent long-range dispersal to new habitats, but inhibiting gene flow between such dispersed populations (Bousquet *et al.*, 1992; Dennis *et al.*, 1995). However, significant results, such as those of Smith (2001), who found a correlation between biotic dispersal and species richness, compared only plants within the same ecological conditions (the tropical understorey).

## 3. Species richness in grasses

The grasses (family Poaceae) are the fifth-largest family in the angiosperms, with almost 10 000 species (Mabberley, 1993). Their importance is beyond doubt, for they provide the grass-dominated ecosystems that cover more than one-third of the Earth's land surface (Archibald, 1995), and they play an essential role in human sustenance, either as a cereal crop or as a source of forage (Raven *et al.*, 1992). The success of the grasses in terms of biodiversity can be explained partly by their adaptability to changeable environments, their ability to resist grazing, and by almost endless variations based on an "all-purpose body plan" (Clayton and Renvoize, 1986; Chapman, 1996).

A great diversity in number of species can be seen between the different grass lineages within the family. For example, four of the ten most important grass subfamilies contain more than 65% of the total number of species (Kellogg, 2000). Of the 395 genera considered in this study, the ten most species-rich represent about one-third of the total grass diversity (3200+

*Table 1.* Character states for the eight phenotypic traits used in the angiosperms and grass analysis. “n/a” means that the data on the particular phenotypic trait was not available for the whole set of taxa considered in this study.

| Trait             | Character state          |            | Supertrees |            |
|-------------------|--------------------------|------------|------------|------------|
|                   | state 1                  | state 2    | grass      | angiosperm |
| Habit             | trees, shrubs and lianas | herbaceous | ✓          | ✓          |
| Life form         | annual and biennial      | perennial  | ✓          | ✓          |
| Sex               | dioecious                | monoecious | ✓          | ✓          |
| Pollination       | wind                     | not wind   | n/a        | ✓          |
| Fruit             | fleshy                   | nonfleshy  | n/a        | ✓          |
| Water requirement | hydrophile               | xerophile  | ✓          | n/a        |
| Salt tolerance    | halophile                | glycophile | ✓          | n/a        |
| Habitat           | open                     | shade      | ✓          | n/a        |

species), and half the total number of species in the family are contained in only 18 genera (Watson and Dallwitz, 1992).

Within the angiosperms, the grasses are thought to be a relatively young family. Although the earliest non-equivocal fossil evidence dates from the early Eocene (~55 million years ago; Crepet and Feldmann, 1991), the global expansion of grasses and their increasing relative abundance in terrestrial ecosystems did not occur before the early to middle Miocene (~15 million years ago; Willis and McElwain, 2002). It has been hypothesized that the co-evolution between grasses and hoofed mammals has had an important role in the expansion of the grasslands and increased speciation within the former (Janis, 1993; MacFadden, 1998). Morphological characters, such as the presence of silica bodies within the leaves, could have had an influence in the success of some species by conferring a higher resistance to herbivory (Chapman, 1996) and allowing open-habitat species to become more species rich. At the same time, compelling evidence suggests that increasing latitudinal aridity promoted the evolution and expansion of grasses (Wing and Boucher, 1998), which could indicate drought resistance as a potential important adaptation in the family. Finally, grasses are part of the angiosperms, and the different traits discussed in Section 2 could have had an influence on grass species richness. In total, we investigated the potential effect of six traits, ranging from anatomical features to life history (Table 1), on the species richness of grasses.

### 3.1 The grass supertree and trait / diversity relationships

The grass supertree was taken from Salamin *et al.* (2002), and is based on 61 published phylogenetic trees that contain a total of 395 grass genera (Figure

1). Because of the ability of supertrees to combine source trees into a more comprehensive tree, almost 50% of all grass genera are represented in this study, which is far greater than the taxonomic coverage offered by molecular phylogenies where typically less than 5% of the genera are represented. The MRP matrix was built using the software SuperTree0.85b (Salamin *et al.*, 2002), and the Baum / Ragan coding scheme was selected with characters weighted by their bootstrap support. The characters of the MRP matrix were considered irreversible during the parsimony analysis (for details, see Salamin *et al.*, 2002). The species number for each genus was recorded from the Grass Genera of the World database (Watson and Dallwitz, 1992; <http://biodiversity.uno.edu/delta/grass/>), and the different morphological characters considered were taken from the Delta database for the grass family (<http://biodiversity.uno.edu/delta/>).

The different traits of interest were mapped on the supertree using ACCTRAN and DELTRAN optimization with the software PAUP\* 4.0b10 (Swofford, 2002). For equivocal character-state reconstructions, ACCTRAN favors reversals of character states over convergences, pushing the origin of the derived character state towards the root of the tree; whereas DELTRAN favors later, parallel origins of the derived character state. Based on these optimizations, all sister clades with contrasting traits were identified. When nested contrasting clades were found, we only selected the most terminal contrasting sister clades; changes from one state to another occurring in the deepest nodes were not considered. Cases where both character states were present within a given taxon could confound the effect of a trait on the species richness of that particular clade. Therefore, we also ignored polymorphic taxa, so that unequivocal contrasts only were examined.

The method of Slowinski and Guyer (1993) was used to compare the number of species belonging to each sister clade against the null hypothesis of equal speciation rates. The method obtains a test statistic from the cumulative probability of obtaining a difference in number of species between the sister clades as large as the one observed. To test the significance of the cumulative probabilities for each of the traits over the different sister clades, we used 1) Fisher's combined probability test as proposed originally by Slowinski and Guyer (1993) and 2) the method of Goudet (1999). Goudet (1999) showed that type I and II errors obtained with Fisher's test can be unduly large as a result of the non-uniform distribution of the probabilities over each sister group, and proposed a randomization procedure to avoid these biases (see also Nee *et al.*, 1996). This procedure is designed to include all cases where the distribution of *p*-values is symmetrical about 0.5 to test the null hypothesis that the distribution of sister-group sizes follows a model of random speciation and extinction. The



Figure 1. Grass supertree, and the position of major clades and subfamilies, based on the Baum / Ragan coding scheme with characters weighted by node support (adapted from Salamin *et al.* (2002)).

*Table 2.* List of traits investigated in the grass family and the results of the tests based on Slowinski and Guyer (1993) and Goudet (1999) methods: a) based on ACCTRAN optimization, and b) based on DELTRAN optimization. States 1 and 2 represent the trait thought to increase species richness and the alternative trait, respectively.

a)

| Trait             | Character states   |            | <i>N</i> | Fisher combined probability test |          | Randomization procedure |          |
|-------------------|--------------------|------------|----------|----------------------------------|----------|-------------------------|----------|
|                   | state 1            | state 2    |          | $\chi^2$                         | <i>p</i> | $G_{\text{obs}}$        | <i>p</i> |
| Life form         | annual             | perennial  | 22       | 96.839                           | <0.001   | 0.258                   | <0.001   |
| Sex               | bisexual / monoecy | dioecy     | 13       | 39.880                           | 0.041    | 0.456                   | 0.342    |
| Habit             | tree-like          | herbaceous | 6        | 38.764                           | 0.015    | 0.316                   | 0.036    |
| Water requirement | hydrophilic        | xerophile  | 20       | 38.804                           | 0.524    | 0.528                   | 0.639    |
| Salt tolerance    | halophile          | glycophile | 9        | 20.048                           | 0.330    | 0.518                   | 0.548    |
| Habitat           | open               | shaded     | 6        | 7.737                            | 0.805    | 0.589                   | 0.752    |

b)

| Trait             | Character states   |            | <i>N</i> | Fisher combined probability test |          | Randomization procedure |          |
|-------------------|--------------------|------------|----------|----------------------------------|----------|-------------------------|----------|
|                   | state 1            | state 2    |          | $\chi^2$                         | <i>p</i> | $G_{\text{obs}}$        | <i>p</i> |
| Life form         | annual             | perennial  | 18       | 67.879                           | <0.001   | 0.304                   | <0.001   |
| Sex               | bisexual / monoecy | dioecy     | 12       | 50.475                           | 0.001    | 0.445                   | 0.192    |
| Habit             | tree-like          | herbaceous | 5        | 34.943                           | <0.001   | 0.298                   | 0.026    |
| Water requirement | hydrophilic        | xerophile  | 21       | 49.454                           | 0.171    | 0.454                   | 0.257    |
| Salt tolerance    | halophile          | glycophile | 9        | 23.287                           | 0.179    | 0.365                   | 0.304    |
| Habitat           | open               | shaded     | 6        | 15.987                           | 0.191    | 0.478                   | 0.543    |

randomization procedure was performed using the software R-1.6.1 (<http://www.r-project.org>). A Bonferroni correction was used because several simultaneous non-independent tests were performed, reducing the alpha value to 0.008 from the nominal 0.05.

### 3.2 Results and discussion

The results showed that the ability to resist drought or salty environments, or to live in open habitat did not affect the speciation rate within the respective grass species possessing those traits significantly (Table 2). These results held regardless of whether ACCTRAN or DELTRAN optimization was used to map the morphological characters onto the supertree. Similarly, being

bisexual or monoecious was not found to have a significant effect using the randomization procedure (Table 2). However, having bisexual spikelets or being a monoecious plant was found to have a significant effect on the number of species with the Fisher's combined probability test under both ACCTRAN and DELTRAN optimization (Table 2). The *p*-values obtained from Fisher's combined probability test were lower in general than those obtained from Goudet's (1999) randomization procedure, confirming the potentially elevated type I error induced by using the former test.

Traits such as drought resistance, salt tolerance, and the ability to thrive in an open habitat are likely to be represented by a broad spectrum of adaptations and phenotypes, and could have evolved from diverse origins. Consequently, the presence of these traits was highly polymorphic within most genera, resulting in the removal of a large number of terminals from the sister-clade comparisons. The decrease in the number of species on either side of the contrasting sister clades arising from the removal of polymorphic taxa can have a large influence on the tests performed, and could have confounded any signs of enhanced diversification rates arising from these traits. The removal of polymorphic taxa could also remove the larger genera preferentially because they are more likely to be polymorphic purely by chance. Ignoring polymorphic taxa could also influence our results by removing important information. For example, the key innovation might actually have an impact on species richness, but would be ignored totally in our comparisons if it arose inside the polymorphic clade. Dioecy has been suggested to be correlated with self-incompatibility, but the interpretation of such a correlation is difficult (Weiblen *et al.*, 2000). However, the link between monoecy and bisexual spikelets and higher diversification rates is done precisely through this assumed correlation. Our results, therefore, could suggest that 1) neither monoecy nor bisexual spikelets correlate with self-compatibility in the grasses because no effect was seen on the rate of diversification, or 2) that the correlation does exist, but that Baker's law is not supported in this family. There was also no significant correlation between a herbaceous habit and an increase in net speciation rate after correction for multiple comparisons (Table 2). In contrast to the five traits presented above, an annual life cycle had a significant effect on species richness in grasses under both tests and regardless of whether ACCTRAN or DELTRAN optimization was used (Table 2). The Fisher's test again gave lower *p*-values than the randomization procedure, which can be explained by the non-uniform distribution of these *p*-values (Goudet, 1999).

Our findings for the grasses support Bousquet *et al.*'s (1992) hypothesis that annuals might be able to fit into new niches better, and, therefore, become more species rich. This hypothesis was based on the faster rates of evolution observed within annuals (Bousquet *et al.*, 1992) and the link

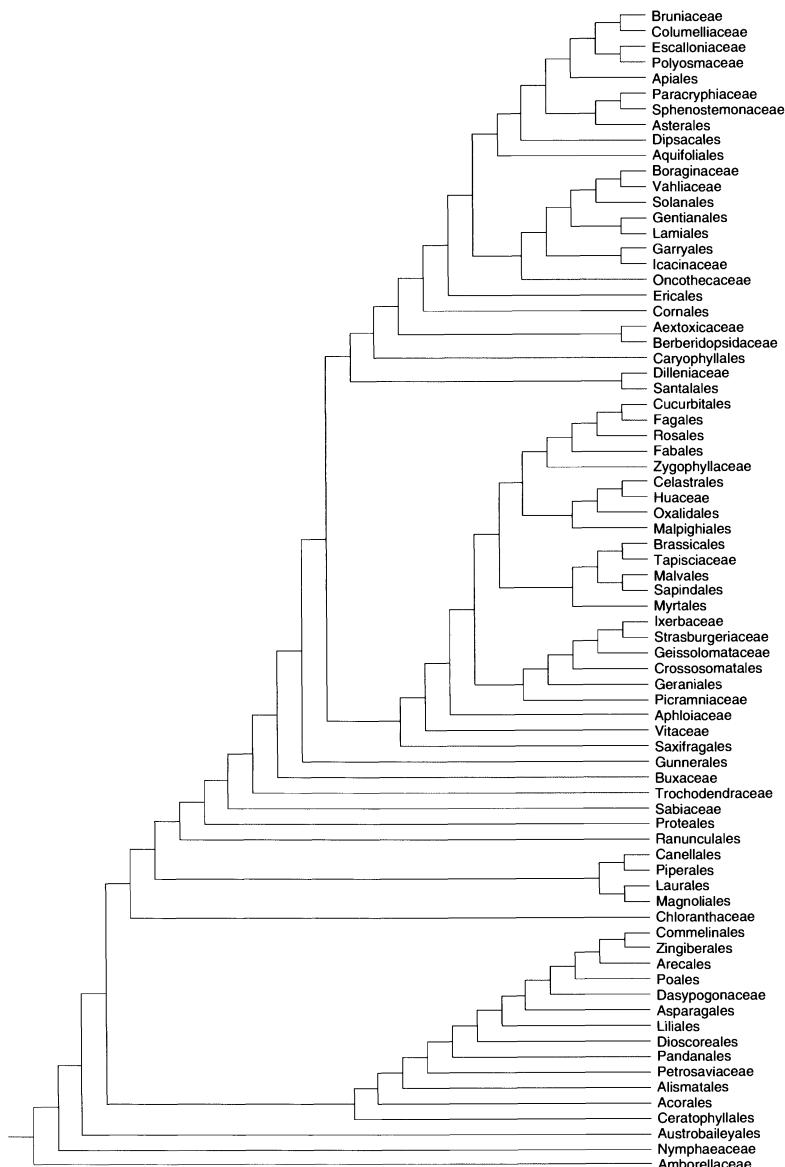
between speciation rates and nucleotide substitution rates (e.g., Barraclough *et al.*, 1996; Savolainen and Goudet, 1998). Although the latter link has been established in other taxonomic groups (Barraclough *et al.*, 1996), evidence for higher speciation rates being correlated with higher substitution rates is inconclusive in the grasses (Gaut *et al.*, 1997). Gaut *et al.*'s (1997) analysis was restricted to a small fraction of grass diversity, and extending the sampling could change the outcome of the analysis. Our results indicate that the generation time, possibly through a change in substitution rate, could influence species richness in the grasses. However, although woody plants such as bamboos have a very long generation time, with some species only flowering every decades (Clayton and Renvoize, 1986), no link between the herbaceous / woody trait and species richness was found. Only a few sister-group comparisons were present on the grass supertree, and most were within the bamboos. Although the paucity of possible sister-group comparisons can have a large effect on the negative results we found, it is probable that factors other than those considered here have also played a role in the success of the grasses.

Another approach to investigate species richness that we did not undertake here is to first identify sister groups with significantly different species richnesses (e.g., using the methods of Moore *et al.*, 2004), and then to look for traits that also differ between these groups. This approach could be more appropriate to highlight whether the possession of a particular phenotypic character was important in the increase of the number of species within a clade. Finally, it has to be noted that no one single trait explains everything, and that the evolutionary responses leading to an increase in species richness are likely to be complex.

#### 4. Species richness in the angiosperms

The flowering plants (angiosperms) represent one of the largest terrestrial radiations, and provide an ideal subject for statistically robust investigations into hypothesized evolutionary explanations for the contemporary pattern in species richness. Over 250 000 species are recognized currently (Wilson, 1992), although estimates vary and the final number might well be double this (Govaerts, 2001; Bramwell, 2002), with familial species richness varying over several orders of magnitude.

The angiosperm supertree (Davies *et al.*, in press; Figure 2) used here is the most complete representation of angiosperm families to date, and allowed us to investigate several alternative hypotheses that have variously been proposed as explaining the diversity of the angiosperms. It is apparent



*Figure 2.* Angiosperm supertree based on the Baum / Ragan coding scheme with characters weighted by node support. Not all families are represented, but ordinal classification is presented when possible (adapted from Davies *et al.*, in press).

from the supertree that some lineages are much more species rich in comparison to their sister lineages and, therefore, appear to have increased rates of diversification. The net result of differential speciation rates is a highly imbalanced topology. Using Purvis *et al.*'s (2001) modification of Fusco and Cronk's (1995) imbalance measure, the weighted mean imbalance was 0.70, which was significantly different ( $p < 0.001$ ) from the expectations of the Markovian null model (Davies *et al.*, *in press*). Although the Markovian null model does not predict a perfectly balanced tree (Kirkpatrick and Slatkin, 1993), the degree of skewness in the imbalance found suggests that the Markovian model is inappropriate for describing the radiation of the angiosperms. There is also evidence that such conclusion holds across a broad taxonomic spectrum (Purvis, 1996; Savolainen *et al.*, 2002).

The use of a supertree approach allowed us, for the first time, to utilize a complete family-level phylogeny of the angiosperms using currently accepted family delineations following the advice of the APG group (Bremer *et al.*, 1998 and onwards). This enabled the most thorough and robust comparisons of clade species richness with regard to putative key traits yet undertaken.

#### 4.1 The angiosperm supertree

The angiosperm supertree combined 46 predominantly molecular phylogenetic trees encompassing the current understanding of angiosperm familial relationships. Source trees were selected based on either their comprehensive coverage or their resolution of relationships that were poorly understood previously (i.e., where support for phylogenetic affinities was weak or absent). In contrast to many situations where supertrees have been used, several large phylogenetic trees of the angiosperms already exist, such as that by Soltis *et al.* (2000) in which around 75% of families are represented. The intention behind constructing the supertree was to amass phylogenetic data to get complete familial representation rather than to produce a consensus of conflicting phylogenetic hypotheses. Consequently, sampling of source trees was not as dense as found in many other studies (e.g., primates, Purvis, 1995; carnivores, Bininda-Emonds *et al.*, 1999; grasses, Salamin *et al.*, 2002). This selection of source trees could leave the supertree construction process open to the accusations of bias, which we argue is not a valid criticism. All phylogenetic analyses suffer from some form of bias in data selection, taxon sampling, or method of analysis (to name but a few); supertree construction is no different in this way. The angiosperm supertree was intended to represent only the most recent understanding of phylogenetic relationships within the group.

Construction of the angiosperm supertree followed broadly that advocated by Salamin *et al.* (2002). As for the grasses, the software SuperTree0.85b (Salamin *et al.*, 2002) using the Baum / Ragan coding scheme was used to build the binary matrix. This matrix was analyzed using parsimony treating all characters as irreversible; further details are given in Davies *et al.* (in press).

## 4.2 Species richness

To investigate correlates of diversity, we used independent contrasts, which is a highly conservative approach with respect to its sensitivity to phylogenetic error (Symonds, 2002). Many previous estimates of phylogeny within the angiosperms are not in complete agreement with one other. However, despite different topologies being generated from different subsets of genes, disagreement between topologies is more likely a product of noise within the data sets rather than an indication of conflicting phylogenetic signal. Consequently, the support for relationships that differ between phylogenetic estimates is low generally. The three genes that have the broadest taxonomic sampling within the angiosperms — *rbcL*, 18S rDNA, and *atpB* — show a high degree of similarity in their phylogenetic signal and general agreement in the relationships they depict among the major angiosperm groups (Soltis *et al.*, 1998). By adopting the protocols of Salamin *et al.* (2002) for weighting relationships within the source trees in the MRP analysis by their respective bootstrap support values, well-supported nodes were able to override less well-supported nodes where there was conflict between the source trees.

Species numbers for sister taxa identified from the angiosperm supertree were obtained from Davies *et al.* (in press) and followed the family delineations of the Angiosperm Phylogeny Group (APG; Bremer *et al.*, 1998; APG II, 2003). When families appeared polyphyletic or paraphyletic in the supertree, they were merged to produce a composite taxon and species richness was calculated as the sum of the number of species in the individual families. States for the five traits mentioned above were also coded using Watson and Dallwitz's (1992) online database. Families recognized currently by the APG, but embedded within larger families in the online database were coded with a question mark. Character states were grouped to correspond to the biological traits under examination, thereby maximizing the number of contrasts. We again ignored polymorphic taxa so that only unequivocal contrasts were examined. We regard this as a conservative test of the key-innovation hypothesis.

For each trait in turn, the character states between the two sister clades were compared at each node on the phylogenetic tree. If these differed, the following species richness contrast was performed:

$$\log \left[ \frac{\text{number of species in the clade possessing state 1}}{\text{number of species in the clade possessing state 2}} \right].$$

A one-sample *t*-test was then performed upon the results for all nodes to see if the mean value differed significantly from zero. A Bonferroni correction was used here as well, reducing the individual alpha value to 0.01 from 0.05.

### 4.3 Results and discussion

One of the most striking aspects of this analysis was the paucity of independent contrasts. Of the five traits examined, two (life form and mode of pollination), produced only two unequivocal sister-taxon comparisons differing in the trait. Interestingly, mode of pollination has been one of the most cited traits in previous attempts to explain the unusual success of the angiosperms. The lack of potential contrasts was as much a consequence of the large number of families that were polymorphic for the trait in question as it was of sister taxa sharing the same trait. Many of the large families such as Asteraceae and Cyperaceae contained species that were both abiotically and biotically pollinated. Those that could be classified easily as one or the other, such as abiotic pollination for the grasses, often shared this trait with their nearest species-poor relatives (Joinvilleaceae and Ecdeiocoleaceae). Of the remaining three comparisons, fruit type produced the most contrasts, but no significant correlation with species richness was apparent ( $p = 0.16$ ; Table 3). A similar lack of significance was found for sex ( $p = 0.98$ ; Table 3) and habit ( $p = 0.62$ ; Table 3).

Several strategies have been adopted in the literature to increase the number of independent contrasts in an effort to test putative key innovations. These methods can be categorized broadly as clade reduction and majority rule. The former approach reduces the species number of a clade by subtracting the number of species that possess the trait deemed atypical of that clade (e.g., Heilbuth, 2000). The latter, and more common, approach characterizes a clade based upon the trait possessed by the majority of species within it (e.g., Eriksson and Bremer, 1992). Both these strategies are unsatisfactory because trait flexibility and species richness are intertwined inextricably (see below). By contrast, the current analysis is the most stringent test of the key-innovation hypothesis.

*Table 3.* List of traits investigated in the angiosperm supertree, number of observations, mean and standard deviation (SD) for each trait, and the one sample *t*-test statistics and associated *p*-value.

| Trait                  | Character state          |                         | <i>N</i> | mean  | SD   | One sample <i>t</i> -test |          |
|------------------------|--------------------------|-------------------------|----------|-------|------|---------------------------|----------|
|                        | state 1                  | state 2                 |          |       |      | <i>t</i>                  | <i>p</i> |
| Habit                  | trees, shrubs and lianas | herbaceous              | 14       | 0.49  | 3.57 | 0.51                      | 0.62     |
| Life Form <sup>1</sup> | annual and biennial      | perennial               | 2        | 1.03  | 1.77 | 0.82                      | 0.56     |
| Sex                    | dioecious                | monoecious              | 10       | -0.02 | 2.32 | -0.03                     | 0.98     |
| Pollination            | wind                     | not wind                | 2        | -4.84 | 3.56 | -1.92                     | 0.31     |
| Fruit                  | fleshy <sup>2</sup>      | non-fleshy <sup>3</sup> | 19       | -0.93 | 2.78 | -1.46                     | 0.16     |

<sup>1</sup> for herbaceous plants

<sup>2,3</sup> as indicators of biotic dispersal

There are many possible reasons why no key innovations were identified in the angiosperms. First, as mentioned above, the species-poor sister clade might also have the key trait associated with increased rates of cladogenesis, but subsequent adaptation in unrelated traits or niche shifts restricted its potential to diversify. Second, different traits might be advantageous at different geological times, with those taxa that happened to be pre-adapted to changes in the environment radiating rapidly. Consequently, particular traits could be correlated with increased rates of diversification only within certain geological time periods. Such a scenario has been suggested as explaining the rapid radiation of the grasses (which had been restricted previously to marginal habitats) coinciding with the late Tertiary change towards a drier climate, which enabled the exploitation of new niches and a dramatic increase in their ecological dominance (Axelrod, 1952; Chapman, 1996). Such an expansion in range size might have also influenced the probability of further speciation by increasing the likelihood of major isolating factors such as geological barriers separating populations (see Rosenzweig, 1992, 1995). The possibility that the rise to dominance of the angiosperms might be as much a consequence of environmental change as a product of evolutionary novelties gains support from the fossil record. The apparent timings of the attainment of dominance varied latitudinally (i.e., was climate specific; Crane and Lidgard, 1989), and the time lag between the origination of particular traits and the apparent increase in the proportion of taxa possessing those traits in the fossil record (Crane *et al.*, 1995) suggests that some factor other than the possession of that particular novel trait was crucial for the subsequent radiations.

If rates of diversification are a product of an interaction between life-history traits and the environment, it might come as no surprise that no

single trait appears to be correlated with contemporary species richness. Over evolutionary time, differing environmental conditions could have favoured the expansion of clades possessing different biological traits, such as biotic pollination in the Orchidaceae around the late Cretaceous (Crane *et al.*, 1995) and abiotic pollination in the grasses in the late Tertiary (Chapman, 1996). Present day species richness is a reflection of the sum of all these historical events, and it might require a unified approach at the interface between knowledge of the fossil record and past climate together with a detailed understanding of phylogeny to tease apart the true story fully. Again, a complementary approach to identifying those traits of importance, suggested by Moore *et al.* (2004), is to identify sister groups with significantly different species richness and to look for traits that differ between them.

An alternative explanation is that traits other than the ones examined might be significant in explaining the success of the angiosperms. Gorelick (2001) lists twenty hypotheses selected from the literature that have variously been proposed to explain the apparent rapid radiation of the angiosperms in comparison to other seed plants, and which encompass co-evolution, breeding system, and numerous other life-history traits. As discussed above, there are certainly many unanswered questions surrounding the potential influence of the traits that we have examined here.

We must of course recognize one further possible explanation for the lack of significance of our findings: that the taxonomic level used in the analysis is inappropriate for identifying correlates of contemporary diversity. This would obviously be the case if the majority of present day species richness was the result of very recent rapid radiations, and if these lineages had not yet achieved sufficient taxonomic distinctiveness to be recognized as separate families. The weakly negative correlation between species richness and family age (Figure 3; see also Burger, 1981) does imply that the majority of the present day species richness could be a product of relatively recent speciation events. However, evidence from the fossil record indicates that the angiosperms attained ecological dominance around 90–130 million years ago (Crane *et al.*, 1995), and that shifts in the rates of diversification within angiosperms have occurred many times over their evolutionary history and across disparate lineages (Davies *et al.*, in press). It is just as probable that both these factors play a part and that contrasting generic-level species richness might give insights into the evolutionary trends favoured by the current environmental conditions, but only a limited understanding of events deeper in time.

Finally, the polymorphic nature of many large families has also led to arguments that it is the very ability to adapt to changing conditions that has enabled some taxa to speciate so rapidly (e.g., Burger, 1981; Rickleff and

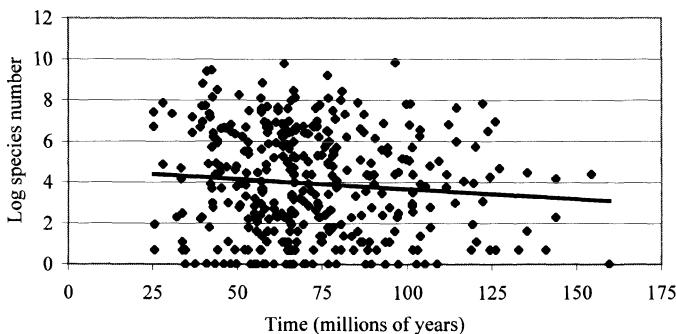


Figure 3. Plot of familial species richness against age of the node from which the family subtends. Dates were taken from Davies *et al.* (in press) and were derived from *rbcL* sequence data obtained from GenBank (<http://www.ncbi.nlm.nih.gov/>). Adjustments for rate heterogeneity among lineages and calibration was performed following Wikström *et al.* (2001).

Renner, 1994). Burger (1981) listed several characters that appear particularly plastic within the angiosperms, including genetic and phenotypic variability in seed production, dispersal and establishment, cell growth, gene expression, and the defining feature of the angiosperms, the flower itself. The extreme plasticity of these traits makes them unsuitable for analysis at the level of all angiosperms. Moreover, a description of these traits across the whole of the angiosperms is not available. A possible hypothesis would be that, rather than possessing a particular character state, being flexible with respect to a given phenotypic trait could lead to an advantage. Given the extraordinary diversity of the angiosperms, such a hypothesis is highly appealing. However, testing such a theory is problematic. The highly variable nature of such traits does not lend itself well to phylogenetic contrasts at higher taxonomic levels such as the family, and characterizing flexibility itself as a trait is beset by an innate circularity. It appears impossible to distinguish whether certain families are more species rich because they have the ability to be flexible in a certain trait or whether it is just more probable that a larger number of character states for a trait will evolve in larger families (see Ricklefs and Renner, 2000; Silvertown *et al.*, 2000). Moreover, all measurements of flexibility are purely inferences drawn from the phylogeny rather than measurable biological characters; and Silvertown *et al.* (2000) argue that, as a consequence, they offer little explanatory power in answering questions about diversification rates. In summary, to gain a better understanding of the causes and processes of diversification, we need an even more detailed knowledge of angiosperm

phylogeny. The current analysis can identify only significant traits that differ between families, and, as observed within the grasses, particular traits might be correlated with species richness within a family, which cannot be tested at this taxonomic level.

## 5. Conclusions

Mixed results were obtained in our investigation of species richness in the grasses and the angiosperms. In the grasses, herbaceous habit and the annual life form were found to be potential key innovations that increased diversification rates. However, for other traits, and also at the higher taxonomic level analyses for the angiosperms, no correlation with species richness was found. It is probable that no simple explanation can offer us a complete understanding of the patterns of contemporary diversity, and a knowledge of evolutionary relationships will become ever more important in providing us with answers to these questions. A goal to aid future investigations would be the creation of a complete generic-level angiosperm phylogeny or a complete species-level grass phylogeny. These are no small tasks, but one that is under consideration currently for the angiosperms. The vast taxonomic sampling required (~10 000 genera and species, respectively) and uncertainties surrounding the limitations of molecular data in resolving such complex phylogenies mean that traditional sequence-based approaches to obtaining these aims are likely to be some way into the future. As illustrated by the examples in this chapter, the use of supertree methodologies such as MRP might make the realization of this objective a much more attainable achievement in the short term.

Our analysis was constrained by several implicit assumptions. Slowinski and Guyer's (1993) method and the species-richness contrasts used here both assume a Markovian model of evolution, which might not be the correct null model of cladogenesis (Cunningham, 1995). At the same time, we can observe only the net speciation rate, and the presence of a clade that is more species rich than its sister counterpart could be a result of either an increase in diversification rates in the larger clade or an increase in extinction rates in the smaller one. Unfortunately, the tests performed here are unable to distinguish between these two cases. Furthermore, the method can be sensitive to errors in the tree used to map the characters of interest and to find contrasting sister clades. As suggested by Dodd *et al.* (1999), it could also be productive to examine interactions between traits or to calculate the variance in diversity associated with different traits that can be obtained with a non-phylogenetic approach (e.g., Ricklefs and Renner, 1994). The development of phylogenetic-based methods coupled with a multivariate

approach could be an extremely useful tool in understanding the origin of differences in species richness between groups of organisms.

Finally, it is important to keep in mind the relative limitation of our approach. The influence of a trait on rates of diversification will probably be contingent upon other taxa, the possession of other traits, and the physical environment. Consequently, no single trait might be associated with species richness at all points on a phylogeny. We therefore advocate that future investigations into patterns of species richness also consider the interactions between biological traits and the environment, as neither is likely to provide definitive answers in isolation. Our goal here was not to look for global answers to the success of the angiosperms or the grasses, but rather to examine whether the possession of a particular character state at a particular point in the phylogeny would be associated with a level of imbalance in the node under consideration.

## Acknowledgements

We would like to thank Olaf Bininda-Emonds for inviting us to contribute to this book; Mark Chase for his invaluable help in ensuring compliance with the APG; and Timothy Barraclough, Trevor Hodkinson, and Vincent Savolainen for their comments and support. We also would like to thank Leslie Watson for providing us with the grass morphological data set, and Elizabeth Zimmer and an anonymous reviewer for comments on the manuscript. This work has been supported by grants from the Swiss National Science Foundation (NS) and a NERC studentship (TJD).

## References

- APG II. 2003. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. *Botanical Journal of the Linnean Society* 141:399–436.
- ARCHIBOLD, O. I. V. 1995. *Ecology of World Vegetation*. Chapman and Hall, London.
- AXELROD, D. I. 1952. A theory of angiosperm evolution. *Evolution* 6:29–60.
- BARRACLOUGH, T. G., NEE, S., AND HARVEY, P. H. 1998. Sister-group analysis in identifying correlates of diversification. *Comment. Evolutionary Ecology* 12:751–754.
- BARRACLOUGH, T. G. AND SAVOLAINEN, V. 2001. Evolutionary rates and species diversity in flowering plants. *Evolution* 55:677–683.
- BAUM, B. R. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon* 41:3–10.
- BAUM, B. R. AND RAGAN, M. A. 2004. The MRP method. In O. R. P. Bininda-Emonds (ed). *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 17–34. Kluwer Academic, Dordrecht, the Netherlands.

- BAWA, K. S. 1994. Pollinators of tropical dioecious angiosperms: a reassessment? No, not yet. *American Journal of Botany* 81:456–460.
- BAWA, K. S. AND OPLER, P. A. 1975. Dioecism in tropical forest trees. *Evolution* 29:167–179.
- BININDA-EMONDS, O. R. P., GITTLEMAN, J. L., AND PURVIS, A. 1999. Building large trees by combining phylogenetic information: a complete phylogeny of the extant Carnivora (Mammalia). *Biological Reviews* 74:143–175.
- BININDA-EMONDS, O. R. P., GITTLEMAN, J. L., AND STEEL, M. A. 2002. The (super)tree of life: procedures, problems, and prospects. *Annual Review of Ecology and Systematics* 33:265–289.
- BOUSQUET, J., STRAUSS, S. H., DOERKSEN, A. H., AND PRICE, R. A. 1992. Extensive variation in evolutionary rate of *rbcL* gene-sequences among seed plants. *Proceedings of the National Academy of Sciences of the United States of America* 89:7844–7848.
- BRAMWELL, D. 2002. How many plant species are there? *Plant Talk* 32:28.
- BREMER K., CHASE M. W., STEVENS P. F., ANDERBERG A. A., BACKLUND A., BREMER B., BRIGGS B. G., ENDRESS P. K., FAY M. F., GOLDBLATT P., GUSTAFSSON M. H. G., HOOT S. B., JUDD W. S., KÄLLERSJÖ M., KELLOGG E. A., KRON K. A., LES D. H., MORTON C. M., NICKRENT D. L., OLSTEAD R. G., PRICE R. A., QUINN C. J., RODMAN J. E., RUDALL P. J., SAVOLAINEN V., SOLTIS D. E., SOLTIS P. S., SYTSMA K. J., AND THULIN M. 1998. An ordinal classification for the families of flowering plants. *Annals of the Missouri Botanic Garden* 85:531–553.
- BURGER, W. C. 1981. Why are there so many kinds of flowering plants. *BioScience* 31:572, 577–581.
- CHAPMAN, G. P. 1996. *The Biology of Grasses*. CAB International, Wallingford, England.
- CHASE, M. W., SOLTIS, D. E., OLSTEAD, R. G., MORGAN, D., LES, D. H., MISHLER, B. D., DUVAL, M. R., PRICE, R. A., HILLS, H. G., QIU, Y.-L., KRON, K. A., RETTIG, J. H., CONTI, E., PALMER, J. D., MANHART, J. R., SYTSMA, K. J., MICHAEL, H. J., KRESS, W. J., KAROL, K. G., CLARK, W. D., HEDREN, M., GAUT, B. S., JANSEN, R. K., KIM, K. J., WIMPEE, C. F., SMITH, J. F., FURNIER, G. R., STRAUSS, S. H., XIANG, Q. Y., PLUNKETT, G. M., SOLTIS, P. S., SWENSEN, S. M., WILLIAMS, S. E., GADEK, P. A., QUINN, C. J., EGUIARTE, L. E., GOLENBERG, E., LEARN, G. H., GRAHAM, S. W., BARRETT, S. C. H., DAYANANDAN, S., AND ALBERT, V. A. 1993 Phylogenetics of seed plants: an analysis of nucleotide-sequence from the plastid gene *rbcL*. *Annals of the Missouri Botanical Garden* 80:528–580.
- CHRISTENSEN, K., DE COLLOBIANO, S. A., HALL, M., AND JENSEN, H. J. 2002. Tangled nature: a model of evolutionary ecology. *Journal of Theoretical Biology* 216:73–84.
- CLAYTON, W. D. AND RENVOIZE, S. A. 1986. *Genera Gramineum, Grass Genera of the World*. Her Majesty's Stationery Office, London.
- CRANE, P. R., FRIS, E. M., AND PEDERSEN, K. J. 1995. The origin and early diversification of angiosperms. *Nature* 374:27–33.
- CRANE, P. R. AND LIDGARD, S. 1989. Angiosperm diversification and paleolatitudinal gradients in cretaceous floristic diversity. *Science* 246:675–246.
- CREPET, W. L. AND FELDMANN, G. D. 1991. The earliest remains of grasses in the fossil record. *American Journal of Botany* 78:1010–1014.
- CUNNINGHAM, S. A. 1995. Problems with null models in the study of phylogenetic radiation. *Evolution* 49:1292–1294.
- DACKS, J. B. AND DOOLITTLE, W. F. 2001. Reconstructing / deconstructing the earliest eukaryotes: how comparative genomics can help. *Cell* 107:419–425.
- DAVIES, T. J., BARRACLOUGH, T. G., CHASE, M. W., SOLTIS, P. S., SOLTIS, D. E. AND SAVOLAINEN, V. In press. Darwin's abominable mystery: insights from a supertree of the

- angiosperms. *Proceedings of the National Academy of Sciences of the United States of America*.
- DENNIS, R. L. H., SHREEVE, T. G., AND WILLIAMS, W. R. 1995. Taxonomic differentiation in species richness gradients among European butterflies (Papilioidea, Hesperioidae): contribution of macroevolutionary dynamics. *Ecography* 18:27–40.
- DODD, M. E., SILVERTOWN, J., AND CHASE, M. W. 1999. Phylogenetic analysis of trait evolution and species diversity variation among angiosperm families. *Evolution* 53:732–744.
- ERIKSSON, O. AND BREMER, B. 1992. Pollination systems, dispersal modes, life forms, and diversification rates in angiosperm families. *Evolution* 46:258–266.
- FARRELL, B. D., DUSSOURD, D. E., AND MITTER, C. 1991. Escalation of plant defense: do latex and resin canals spur plant diversification. *American Naturalist* 138:881–900.
- FELSENSTEIN, J. 1985. Phylogenies and the comparative method. *American Naturalist* 125:1–15.
- FOLEY, R. 2002. Adaptive radiations and dispersals in hominid evolutionary ecology. *Evolutionary Anthropology* 11:32–37.
- FUSCO, G. AND CRONK, Q. C. B. 1995. A new method for evaluating the shape of large phylogenies. *Journal of Theoretical Biology* 175: 235–243.
- GATESY, J., MATTHEE, C., DESALLE, R., AND HAYASHI, C. 2002. Resolution of a supertree / supermatrix paradox. *Systematic Biology* 51:652–664.
- GAUT, B. S., CLARK, L. G., WENDEL, J. F., AND MUSE, S. V. 1997. Comparisons of the molecular evolutionary process at *rbcL* and *ndhF* in the grass family (Poaceae). *Molecular Biology and Evolution* 14:769–777.
- GAUT, B. S., MORTON, B. R., MCCAIIG, B. C., AND CLEGG, M. T. 1996. Substitution rate comparisons between grasses and palms: Synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcL*. *Proceedings of the National Academy of Sciences of the United States of America* 93:10274–10279.
- GAUT, B. S., MUSE, S. V., CLARK, W. D., AND CLEGG, M. T. 1992. Relative rates of nucleotide substitution at the *rbcL* locus of monocotyledonous plants. *Journal of Molecular Evolution* 35:292–303.
- GITTLEMAN, J. L., JONES, K. E., AND PRICE, S. A. 2004. Supertrees: using complete phylogenies in comparative biology. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 439–460. Kluwer Academic, Dordrecht, the Netherlands.
- GITTLEMAN, J. L. AND PURVIS, A. 1998. Body size and species-richness in carnivores and primates. *Proceedings of the Royal Society of London B*. 265:113–119.
- GORELIK, R. 2001. Did insect pollination cause increased seed plant diversity? *Biological Journal of the Linnean Society* 74:407–427.
- GOUDET, J. 1999. An improved procedure for testing the effects of key innovations on rate of speciation. *American Naturalist* 153:549–555.
- GOVAERTS, R. 2001. How many species of seed plants are there? *Taxon* 50:1085–1090.
- GRAYBEAL, A. 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? *Systematic Biology* 47:9–17.
- HALANYCH, K. M. AND PASSAMANECK, Y. 2001. A brief review of metazoan phylogeny and future prospects in Hox-research. *American Zoologist* 41:629–639.
- HEILBUTH, J. C. 2000. Lower species richness in dioecious clades. *American Naturalist* 156:221–241.
- HILLIS, D. M. 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. *Systematic Biology* 47:1–8.

- JANIS, C. M. 1993. Tertiary mammal evolution in the context of changing climates, vegetation, and tectonic events. *Annual Review of Ecology and Systematics* 24:467–500.
- JEFFERY, J. E., RICHARDSON, M. K., COATES, M. I., AND BININDA-EMONDS, O. R. P. 2002. Analyzing developmental sequences within a phylogenetic framework. *Systematic Biology* 51:478–491.
- JOHNSON, C. N. 1998. Species extinction and the relationship between distribution and abundance. *Nature* 394:272–274.
- KÄLLERSJÖ, M., FARRIS, J. S., CHASE, M. W., BREMER, B., FAY, M. F., HUMPHRIES, C. J., PETERSEN, G., SEBERG, O., AND BREMER, K. 1998. Simultaneous parsimony jackknife analysis of 2538 *rbcL* DNA sequences reveals support for major clades of green plants, land plants, seed plants and flowering plants. *Plant Systematics and Evolution* 213:259–287.
- KELLOGG, E. A. 2000. The grasses: a case study in macroevolution. *Annual Review of Ecology and Systematics* 31:217–238.
- KIRKPATRICK, M. AND SLATKIN, M. 1993. Searching for evolutionary patterns in the shape of a phylogenetic tree. *Evolution* 47:1171–1181.
- KOCH, M. A., WEISSHAAR, B., KROYMANN, J., HAUBOLD, B., AND MITCHELL-OLDS, T. 2001. Comparative genomics and regulatory evolution: conservation and function of the *Chs* and *Apetala3* promoters. *Molecular Biology and Evolution* 18:1882–1891.
- LARCHER, W. 1995. *Physiological Plant Ecology: Ecophysiology of Functional Groups*. Springer-Verlag, Berlin.
- LINDER, H. P. 2000. Vicariance, climate change, anatomy and phylogeny of Restionaceae. *Botanical Journal of the Linnean Society* 134:159–177.
- MABBREY, D. 1993. *The Plant-Book: a Portable Dictionary of the Vascular Plants*. Cambridge University Press, Cambridge.
- MACFADDEN, B. J. 1998. Tale of two rhinos: isotopic ecology, paleodiet, and niche differentiation of *Aphelops* and *Teleoceras* from the Florida Neogene. *Paleobiology* 24:274–286.
- MANNING, J. C. AND LINDER, H. P. 1992. Pollinators and evolution in *Disperis* (Orchidaceae), or why are there so many species. *South African Journal of Science* 88:38–49.
- MARVALDI, A. E., SEQUEIRA, A. S., O'BRIEN, C. W., AND FARRELL, B. D. 2002. Molecular and morphological phylogenetics of weevils (Coleoptera, Curculionoidea): do niche shifts accompany diversification? *Systematic Biology* 51:761–785.
- MARZLUFF, J. M. AND DIAL, K. P. 1991. Life history correlates of taxonomic diversity. *Ecology* 72:428–439.
- MAYNARD SMITH, J. AND SZATHMARY, E. 1995. *The Major Transitions in Evolution*. Freeman, Oxford.
- MIDGELEY, J. J. AND BOND, W. J. 1991. How important is biotic pollination and dispersal to the success of the angiosperms? *Philosophical Transactions of the Royal Society of London B* 333:209–215.
- MOOERS, A. Ø. AND HARVEY, P. H. 1994. Metabolic rate, generation time, and the rate of molecular evolution in birds. *Molecular Phylogenetics and Evolution* 3:344–350.
- MOORE, B. R., CHAN, K. M. A., AND DONOGHUE, M. J. 2004. Detecting diversification rate variation in supertrees. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 487–533. Kluwer Academic, Dordrecht, the Netherlands.
- NUNN, C. L. AND BARTON, R. A. 2000. Allometric slopes and independent contrasts: a comparative test of Kleiber's law in primate ranging patterns. *American Naturalist* 156:519–533.

- PARADIS, E. 1998. Detecting shifts in diversification rates without fossils. *American Naturalist* 152:176–188.
- PURVIS, A. 1995. A composite estimate of primate phylogeny. *Philosophical Transactions of the Royal Society of London B* 348:405–421.
- PURVIS, A. 1996. Using interspecies phylogenies to test macroevolutionary hypotheses. In P. H. Harvey, A. J. Leigh Brown, J. Maynard Smith, and S. Nee (eds), *New Uses for New Phylogenies*, pp.153–168. Oxford University Press, Oxford.
- PURVIS, A., KATZOURAKIS, A., AND AGAPOW, P.-M. 2001. Evaluating phylogenetic tree shape: two modifications to Fusco and Cronk's method. *Journal of Theoretical Biology* 214:99–103.
- PURVIS, A., NEE, S., AND HARVEY, P. H. 1995. Macroevolutionary inferences from primate phylogeny. *Proceedings of the Royal Society of London B* 260:329–333.
- QIU, Y.-L., LEE, J., BERNASCONI-QUADRONE, F., SOLTIS, D. E., SOLTIS, P. S., ZANIS, M., CHEN, Z., SAVOLAINEN, V., AND CHASE, M. W. 2000. Phylogeny of basal angiosperms: analysis of five genes from three genomes. *International Journal of Plant Sciences* 161:S3–S27.
- RAGAN, M. A. 1992. Phylogenetic inference based on matrix representation of trees. *Molecular Phylogenetics and Evolution* 1:53–58.
- RAVEN, P. H., EVERET, R. H., AND EICHHORN, S. E. 1992. *Biology of Plants*. Worth Publishers, New York.
- RICKLEFS, R. E. AND RENNER, S. S. 1994. Species richness within families of flowering plants. *Evolution* 48:1619–1636.
- RIESEBERG, L. H. 1997. Hybrid origins of plant species. *Annual Review of Ecology and Systematics* 28:359–389.
- ROSENHEIM, J. A. AND TABASHNIK, B. E. 1991. Influence of generation time on the rate of response to selection. *American Naturalist* 137:527–541.
- ROSENZWEIG, M. L. 1992. Species diversity gradients: we know more and less than we thought. *Journal of Mammalogy* 73:715–730.
- SALAMIN, N., HODKINSON, T. R., AND SAVOLAINEN, V. 2002. Building supertrees: an empirical assessment using the grass family (Poaceae). *Systematic Biology* 51:136–150.
- SANDERSON, M. J. AND DONOGHUE, M. J. 1994. Shifts in diversification rate with the origin of angiosperms. *Science* 264:1590–1593.
- SAVOLAINEN, V., HEARD, S. B., POWELL, M., DAVIES, T. J., AND MOOERS, A. Ø. 2002. Is cladogenesis heritable? *Systematic Biology* 51:1–9.
- SAVOLAINEN, V., CHASE, M. W., MORTON, C. M., HOOT, S. B., SOLTIS, D. E., BAYER, C., FAY, M. F., DEBRUIJN, A., SULLIVAN, S., AND QIU, Y.-L. 2000. Phylogenetics of flowering plants based upon a combined analysis of plastid *atpB* and *rbcL* gene sequences. *Systematic Biology* 49:306–362.
- SAVOLAINEN, V. AND GOUDET, J. 1998. Rate of gene sequence evolution and species diversification in flowering plants: a re-evaluation. *Proceedings of the Royal Society of London B* 265:603–607.
- SCHMID-HEMPPEL, P. AND EBERT, D. 2003. On the evolutionary ecology of specific immune defence. *Trends in Ecology and Evolution* 18:27–32.
- SEMPLE, C. AND STEEL, M. A. 2000. A supertree method for rooted trees. *Discrete and Applied Mathematics* 105:147–158.
- SILVERTOWN, J., MC CONWAY, K. J., DODD, M. E., AND CHASE, M. W. 2000. “Flexibility” as a trait and methodological issues in species diversity variation among angiosperm families. *Evolution* 54:1066–1068.

- SIMPSON, P. 2002. Evolution of development in closely related species of flies and worms. *Nature Reviews Genetics* 3:907–917.
- SLOWINSKI, J. B. AND GUYER, C. G. 1993. Testing whether certain traits have caused amplified diversification: an improved method based on a model of random speciation and extinction. *American Naturalist* 142:1019–1024.
- SMITH, J. F. 2001. High species diversity in fleshy-fruited tropical understory plants. *American Naturalist* 157:646–653.
- SOLTIS, D. E., SOLTIS, P. S., ALBERT, V. A., OPPENHEIMER, D. G., DE PAMPHILIS, C. W., MA, H., FROHLICH, M. W., AND THEISSEN, G. 2002. Missing links: the genetic architecture of flower and floral diversification. *Trends in Plant Science* 7:22–31.
- SOLTIS, D. E., SOLTIS, P. S., CHASE, M. W., MORT, M. E., ALBACH, D. C., ZANIS, M., SAVOLAINEN, V., HAHN, W. H., HOOT, S. B., FAY, M. F., AXTELL, M., SWENSEN, S. M., PRINCE, L. M., KRESS, W. J., NIXON, K. C., AND FARRIS, J. S. 2000. Angiosperm phylogeny inferred from 18S rDNA, *rbcL*, and *atpB* sequences. *Botanical Journal of the Linnean Society* 133:381–461.
- SOLTIS, D. E., SOLTIS, P. S., MORT, M. E., CHASE, M. W., SAVOLAINEN, V., HOOT, S. B., AND MORTON, C. M. 1998. Inferring complex phylogenies using parsimony: an empirical approach using three large DNA data sets for angiosperms. *Systematic Biology* 47:32–42.
- SOLTIS, P. S., SOLTIS, D. E., WOLF, P. G., NICKRENT, D. L., CHAW, S., AND CHAPMAN, R. L. 1999. The phylogeny of land plants inferred from 18S rDNA sequences: pushing the limits of rDNA signal? *Molecular Biology and Evolution* 16:1774–1784.
- SWOFFORD, D. L. 2002. *PAUP\*. Phylogenetic Analysis Using Parsimony (\*and Other Methods)*. Version 4. Sinauer, Sunderland, Massachusetts.
- SYMONDS, M. R. E. 2002. The effect of topological inaccuracy in evolutionary trees on the phylogenetic comparative method of independent contrasts. *Systematic Biology* 51:541–553.
- WATSON, L. AND DALLWITZ, M. J. 1992. *The Grass Genera of the World*. CAB International, Wallingford, England.
- WIKSTRÖM, N., SAVOLAINEN, V., AND CHASE, M. W. 2001. Evolution of the angiosperms: calibrating the family tree. *Proceedings of the Royal Society of London B* 268:2211–2220.
- WILKINSON, M., THORLEY, J. L., LITTLEWOOD, D. T. J., AND BRAY, R. A. 2001. Towards a phylogenetic supertree for the Platyhelminthes? In D. T. J. Littlewood and R. A. Bray (eds), *Interrelationships of the Platyhelminthes*, pp. 292–301. Chapman-Hall, London.
- WILKINSON, M., THORLEY, J. L., PISANI, D., LAPointe, F.J., AND MCINERNEY, J. O. 2004. Some desiderata for liberal supertrees. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 227–246. Kluwer Academic, Dordrecht, the Netherlands.
- WILLIS, K. AND MCÉLWAIN, J. 2002. *The Evolution of Plants*. Oxford University Press, Oxford.
- WING, S. L. AND BOUCHER, L. D. 1998. Ecological aspects of the Cretaceous flowering plant radiation. *Annual Review of Earth and Planetary Sciences* 26:379–421.
- ZWICKL, D. J. AND HILLIS, D. M. 2002. Increased taxon sampling greatly reduces phylogenetic error. *Systematic Biology* 51:588–598.

## Chapter 22

# DETECTING DIVERSIFICATION RATE VARIATION IN SUPERTREES

Brian R. Moore, Kai M. A. Chan, and Michael J. Donoghue

**Abstract:** Although they typically do not provide reliable information on divergence times, supertrees are nevertheless attractive candidates for the study of diversification rates: by combining a collection of less inclusive source trees, they promise to increase both the number and density of taxa included in the composite phylogeny. The relatively large size and possibly more dense taxonomic sampling of supertrees have the potential to increase the statistical power and decrease the bias, respectively, of methods for studying diversification rates that are robust to uncertainty regarding the timing of diversification events. These considerations motivate the development of atemporal methods that can take advantage of recent and anticipated advances in supertree estimation. Herein, we describe a set of whole-tree, topology-based methods intended to address two questions pertaining to the study of diversification rates. First, has a given (super)tree experienced significant variation in diversification rates among its branches? Second, if so, where have significant shifts in diversification rate occurred? We present results of simulation studies that characterize the statistical behavior of these methods, illustrating their increased power and decreased bias. We also applied the methods to a published supertree of primates, demonstrating their ability to contend with relatively large, incompletely resolved (super)trees. All the methods described in this chapter have been implemented in the freely available program, SYMMETREE.

**Keywords:** cladogenesis; diversification rate shifts; diversification rate variation; equal-rates Markov random branching model; extinction; Primates; speciation; supertrees; tree shape; Yule branching process

## 1. Introduction

Supertrees represent somewhat of a mixed bag for the study of diversification rates, providing some kinds of information in unprecedented profusion but inherently limited in their ability to provide other types of pertinent data. Ideally, (super)trees can provide two sources of information relevant to the study of diversification rates: the temporal distribution of branching events through time and the topological distribution of species diversity across its branches<sup>1</sup>.

It is generally accepted that, by virtue of directly incorporating information on the timing of diversification, temporal methods enjoy an advantage in power relative to their topological counterparts (e.g., Sanderson and Donoghue, 1996; Paradis, 1998a, b). This power advantage has, in turn, motivated the elaboration of temporal methods to effectively address a relatively wide range of evolutionary questions related to diversification rates. Unfortunately, existing supertree methods typically do not provide reliable branch-length estimates (but note recent progress by Lapointe and Cucumel, 1997; Bryant *et al.*, 2004; Lapointe and Levasseur, 2004; Vos and Mooers, 2004), essentially precluding the use of more powerful temporal methods for the inference of diversification rates.

On the other hand, any decrease in power associated with the necessary reliance on topological methods might be offset to some extent by the typically larger size of supertrees because the power of these methods is known to scale with tree size (e.g., Kirkpatrick and Slatkin, 1993; Kubo and Iwasa, 1995; Paradis, 1997, 1998a, b; Agapow and Purvis, 2002). More than just their potentially larger size, however, is the promise of supertrees to greatly increase the density of sampled taxa. Both temporal and topological methods are sensitive to incomplete and/or nonrandom taxon sampling (e.g., Kubo and Iwasa, 1995; Nee *et al.*, 1996; Pybus and Harvey, 2000; Barraclough and Nee, 2001) for the simple reason that these methods do not discriminate between species that have been omitted from a phylogenetic analysis and those that have been eliminated by extinction. The relatively broad and dense taxonomic sampling of supertrees should therefore confer

<sup>1</sup> Two corresponding classes of methods have been developed to exploit these different sources of information (Sanderson and Donoghue, 1996). The first class relies exclusively on topological information, comparing the observed difference in species diversity between two (or more) groups descended from a common node to the expectation generated under a stochastic model of diversification (e.g., Slowinski and Guyer, 1989a, b, 1993; Slowinski, 1990). The second class utilizes estimates of branch length or duration to infer the (absolute or relative) timing of speciation events and similarly compares the observed distribution of speciation events through time with that expected under a null model of random diversification (e.g., Harvey *et al.*, 1991, 1994a, b; Hey, 1992; Nee *et al.*, 1992, 1994a, b, 1995, 1996; Harvey and Nee, 1993, 1994; Sanderson and Bharathan, 1993; Kubo and Iwasa, 1995; Paradis, 1997, 1998a, b; Pybus and Harvey, 2000; Nee, 2001; Pybus *et al.*, 2002). We refer to these two approaches as *topological* and *temporal* methods, respectively (Chan and Moore, 2002).

increased statistical power and decreased bias to studies of diversification rates, which motivates the development of methods that do not rely on temporal information.

Furthermore, even when reliable branch-length estimates are available, there might be situations in which it is preferable to omit these data from studies of diversification rates. Several types of evolutionary study entail hypothesized associations (whether correlational or causal in nature) between diversification rates and some other variable that is conditioned on branch lengths/durations. For example, there is considerable interest in exploring the putative correlation between rates of diversification and rates of molecular evolution (e.g., Mindell *et al.*, 1989; Barraclough *et al.*, 1996; Savolainen and Goudet, 1998; Barraclough and Savolainen, 2001; Jobson and Albert, 2002). Similarly, many evolutionary questions pertain to the relationship between rates of diversification and rates (and/or ancestral states) of morphological evolution. Often, rate estimates for such variables are either directly or indirectly conditioned on branch-length estimates (e.g., model-based inference of rates of nucleotide substitution, and model-based inference of rates and/or ancestral states of morphological character evolution, respectively). Consequently, attempts to understand the correlation of such variables to variation in rates of diversification will be confounded if both are conditioned on the same set of branch-length estimates. For such inference problems, it would therefore be desirable to possess methods that do not rely on branch-length data.

Accordingly, the nature of the data at hand and/or the hypotheses of interest will often preclude the inference of diversification rates based on temporal information. Clearly, topological methods warrant further consideration. In this chapter, we extend existing topological methods in new ways to exploit new opportunities. Because different people have different interests in the study of differential diversification rates, we describe a suite of methods intended to address two different questions: 1) has a given tree experienced significant variation in diversification rates among its branches; and, 2) if so, on which branches have significant shifts in diversification rate occurred? We explore the statistical behavior of the various methods by means of simulation and illustrate their application to empirical data using a published supertree of primates (Purvis, 1995). Choice of this data set was motivated by two considerations: the primate supertree is in many respects representative of those published for other groups (e.g., in its size, degree of resolution, and methods of estimation), and this tree has been used previously to explore various aspects of diversification rates in primates (e.g., Purvis *et al.*, 1995), thereby affording comparison of our results to those derived with other methods. All the methods described in this chapter have been implemented in the freely

available software program, SYMMETREE (<http://www.kchan.org> or <http://www.phylogenetics.net/brian/>).

## 2. The equal-rates Markov random branching model

The ability of phylogenies to inform studies of differential diversification rates has been appreciated for some time. Hennig (1966) reasoned that any difference in species diversity between two sister groups, which are by definition of equal age, must necessarily reflect different rates of diversification (i.e., speciation minus extinction) in those groups. However, other researchers were quick to caution against overly deterministic interpretations of such differences: even if the underlying probability of diversification were identical in all lineages, some degree of variation in their realized diversification rates would be expected to arise because of the inherently stochastic nature of the branching process (e.g., Raup *et al.*, 1973; Gould *et al.*, 1977).

In recognition of the nature of the process under study, stochastic branching process models are frequently employed to generate an expected distribution of differences in diversity against which observed differences can be compared. One of the most elemental and frequently invoked models is the so-called equal-rates Markov (ERM) random branching process (Yule, 1924; Kendall, 1948; Harding, 1971). This is a continuous-time, discrete-state, pure-birth Markov process in which the probability of a branching event,  $\lambda$ , is constant for each tip in a growing tree at any moment in time<sup>2</sup>. Under the ERM model, the allocation of diversity among two sister groups follows a uniform distribution, such that all possible partitions of  $N$  species,  $1:(N - 1)$ ,  $2:(N - 2)$ ,  $3:(N - 3)$  ...  $(N - 1):1$ , are equiprobable. Accordingly, given an observed diversity partition of  $N$  into  $\ell$  and  $r$  species among two sister groups, we can calculate the cumulative probability of realizing a diversity partition as or more extreme under the ERM model as

$$(1) \quad P = \frac{2\ell}{\binom{N-1}{\ell}}$$

<sup>2</sup> Note that the ERM model allows  $\lambda$  to vary through time, so long as it is equal across all tips at any instant (e.g., Harding, 1971). This property of the ERM model technically distinguishes it from the more restricted constant-rate, pure-birth Yule branching process model because the latter constrains  $\lambda$  to be constant both across tips and through time (e.g., Yule, 1924). Nevertheless, the two models are operationally identical when branching times are unknown, as is the case for topology-based inferences of diversification rate.

(unless  $\ell = N / 2$ , in which case  $P = 1$ ), where  $\ell$  is the number of species in the less diverse of the two sister groups (Slowinski and Guyer, 1989a). A significant difference in sister-group diversity constitutes rejection of the ERM null model, and therefore, suggests that the two lineages have diversified under significantly different rates (Slowinski and Guyer, 1989a, b; Slowinski, 1990). For convenience, we refer to these  $P$ -values as *ERM nodal probabilities* because they pertain to the cumulative ERM probability of realizing a diversity partition between lineages descended from a shared node.

Derivation of an ERM nodal probability incorporates minimal information on the topological distribution of species diversity (only two observations are made). Because the statistical power of a test is a function of sample size, the sensitivity of these single-node tests to differential diversification rates is quite low (e.g., Kirkpatrick and Slatkin, 1993; Fusco and Cronk, 1995; Sanderson and Donoghue, 1996; Sanderson and Wojciechowski, 1996). As we will demonstrate in the following sections, however, these nodal probabilities can serve as building blocks that can be variously generalized to construct methods that harness their collective power.

### 3. Detecting among-lineage diversification rate variation

In this section we consider the question, “Has a given tree experienced significant diversification rate variation among its branches?” This is the diversification rate analogue to the problem of detecting among-lineage substitution rate variation in studies of molecular evolution. The ability to detect among-lineage diversification rate variation has parallel applications to tests of the molecular clock: tests of rate homogeneity are a prerequisite for the application of several temporal methods that assume negligible levels of among-lineage diversification rate variation (e.g., Hey, 1992; Harvey *et al.*, 1994a, b; Nee *et al.*, 1994a, b; Kubo and Iwasa, 1995; Paradis, 1997, 1998a, b; Pybus and Harvey, 2000). Additionally, and like its molecular counterpart, the study of diversification rate variation has important evolutionary implications that might be of interest in their own right (Chan and Moore, 2002).

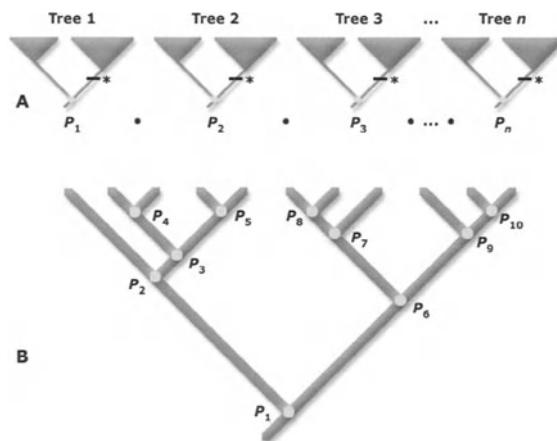
Previous work on this problem has largely involved the development of “tree-balance indices”, metrics that variously summarize the topological distribution of species diversity as a single number. Approximately 20 such indices have been proposed (e.g., Colless, 1982; Shao and Sokal, 1990; Heard, 1992; Kirkpatrick and Slatkin, 1993; Page, 1993; Fusco and Cronk,

1995; McKenzie and Steel, 2000; Agapow and Purvis, 2002; Purvis *et al.*, 2002). Several authors have noted that these indices appear to capture different but poorly characterized aspects of tree shape (Shao and Sokal, 1990; Kirkpatrick and Slatkin, 1993; Fusco and Cronk, 1995). Consequently, all attempts to test for significant diversification rate variation with these tree-balance indices must grapple with the “agony of choice” between myriad alternatives or opt to use all (or some subset of) the indices and endure issues of multiple-test correction. In any case, interpretation of results under the chosen index (or indices) is apt to be less than straightforward: these indices are not derived explicitly from any model of diversification, such that the biological meaning of “significant imbalance” under these tests is unclear.

Our approach to the problem draws on the analogy to the study of among-lineage substitution rate variation: just as single-node tests (as implemented by the relative-rate test; e.g., Sarich and Wilson, 1967; Wu and Li, 1985) have been variously generalized over the whole tree (e.g., Felsenstein 1988, 1989; Takezaki *et al.*, 1995) to realize substantially increased sensitivity to substitution-rate variation, our strategy is to generalize single-node tests (as implemented by ERM nodal probabilities) over the whole tree with the similar objective of increasing the power to detect diversification rate variation. Our presentation of these whole-tree methods necessarily draws upon our previous work (Chan and Moore, 2002) but includes several new results, including the development of two new statistics and a simulation-based exploration of their statistical behavior.

### 3.1 Whole-tree tests of diversification rate variation

Generalization of the single-node approach to incorporate information on the relative diversity of all internal nodes of a tree would provide a much more powerful and — by virtue of being based on an explicit model of cladogenesis — also biologically meaningful test of among-lineage diversification rate variation. The development of such whole-tree methods might be achieved by combining individual ERM nodal probabilities on a node-by-node basis over all internal nodes of a given phylogeny (J. Slowinski, pers. comm. to Kirkpatrick and Slatkin, 1993). But how should individual nodal probabilities be combined? A subsequent development by Slowinski and Guyer (1993) suggests a possible solution. They proposed a method for combining individual ERM probabilities from single-node comparisons from many *different* trees using Fisher’s combined probability test (FCPT; Fisher, 1932). It would seem relatively straightforward to modify the FCPT protocol to combine probabilities from many nodes within the *same* tree (Figure 1).



*Figure 1.* Combining nodal probabilities to develop whole-tree tests of diversification-rate variation. A) Slowinski and Guyer (1993) proposed combining individual ERM nodal probabilities (derived using equation (1)), each from a different tree, using Fisher's combined probability test (FCPT) to test the cumulative effect of a putative key innovation on rates of diversification in the various groups in which it evolved independently (indicated by asterisks). B) Whole-tree tests of diversification-rate variation could seemingly be developed by using FCPT (or ECPT) to combine the individual ERM nodal probabilities from many nodes within the same tree (e.g.,  $P_1$ – $P_{10}$ ). However, the FCPT and ECPT tests assume that the individual probabilities to be combined are independent and can each realize any value between 0 and 1. Nodal probabilities, however, are both non-independent (e.g.,  $P_4$  and  $P_5$  are nested phylogenetically within  $P_3$ ) and valued discretely (they are derived from the comparison of discretely valued species numbers). Nevertheless, approximate solutions can be devised that allow for the combination of nodal probabilities by using Monte Carlo simulation to estimate the appropriate distribution of the test statistics.

Although intuitively appealing, the combination of nodal probabilities under the FCPT is extremely biased. This bias stems from violation of the underlying assumptions of omnibus statistics (i.e., statistics that, like the FCPT, reflect the combined significance of several independent tests of a common hypothesis). The FCPT statistic is calculated by estimating the compound probability that a set of probabilities (in this case, the set of ERM nodal probabilities derived with equation (1)) has a product equal to or smaller than that of the observed set (Fisher, 1932). A less common but equally valid omnibus statistic proposed by Edgington (ECPT: 1972a, b) takes the sum rather than the product of individual probabilities. Both the FCPT and ECPT assume that the individual probabilities to be combined are independent and can realize any value on the interval (0, 1]. However, nodal probabilities are interdependent to the extent that they are derived from phylogenetically nested nodes and these probabilities can realize only a

finite number of discrete values for the simple reason that they are derived (using equation (1)) from the comparison of species diversities, which necessarily occur as whole numbers (i.e., 1, 2, 3, ...). This “discreteness” problem is known to cause a discrepancy between the assumed and realizable probability space (Wallis, 1942; Edgington and Haller, 1984), such that the combination of individual nodal probabilities under the FCPT or ECPT will assume a concave function of the true cumulative probabilities.

In view of the complications associated with the use of conventional omnibus statistics for this problem, we pursue a non-analytical solution that avoids the discreteness and interdependence problems while emulating the logic of the FCPT and ECPT statistics. We first review two whole-tree tests of diversification rate variation based on the cumulative ERM probability derived from the product ( $M_{\Pi}$ ) and sum ( $M_{\Sigma}$ ) of individual nodal probabilities (Chan and Moore, 2002) and then develop two modified versions of these whole-tree statistics,  $M_{\Pi}^*$  and  $M_{\Sigma}^*$ , that differentially weight the individual ERM nodal probabilities according to their species diversity. Conceptually, these four tests involve mapping the sample space that can be realized by discretely valued, interdependent ERM nodal probabilities. This entails the use of Monte Carlo simulation to estimate the underlying distribution of topologies that can be realized for a tree of a given size.

These tests are implemented with one of two algorithms depending upon the size of the tree in question. For smaller trees ( $N < 20$ ), the appropriate ERM sample space can be mapped exactly by applying the “small-tree” algorithm as follows: 1) Calculate the product (or sum) of all ERM nodal probabilities (derived by equation (1)) in the observed tree. 2) Generate all possible topologies for a tree with the same number of species as the observed tree. For each topology, calculate the product (or sum) of its nodal probabilities and its point probability under the ERM model. 3) Sum the point probabilities of all topologies with nodal probability products (or sums) less than or equal to that of the observed tree. This sum represents the cumulative whole-tree probability based on the nodal probability product,  $M_{\Pi}$  (or on the nodal probability sum,  $M_{\Sigma}$ ).

For larger trees ( $N > 20$ ), the appropriate ERM sample space must be approximated using the “large-tree” algorithm owing to the vast number of possible topologies (e.g., only 46 for nine species, but 105 061 603 969 for 35 species; Stone and Repka, 1998). The large-tree algorithm is executed as follows: 1) As in the small-tree algorithm, first calculate the product (or sum) of ERM nodal probabilities in the observed tree. 2) Using the ERM model of cladogenesis, generate a large, random subset of possible topologies for a tree with the same number of species as the observed tree. 3) Count the number of simulated trees with a nodal probability product (or

sum) less than or equal to that of the observed tree and divide by the total number of simulated trees. This quotient is an unbiased estimate of the probability corresponding to  $M_{\Pi}$  (or  $M_{\Sigma}$ ).

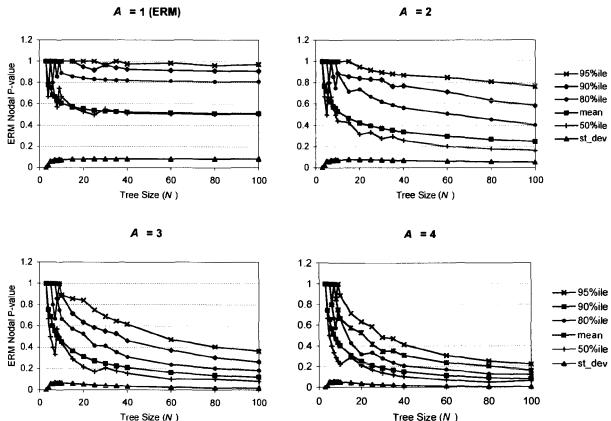
Note that all nodal probabilities contribute equally to the calculation of the  $M_{\Pi}$  and  $M_{\Sigma}$  whole-tree statistics. However, larger nodes (i.e., those defining more diverse clades) sample a greater number of diversification events and should, therefore, provide more reliable evidence of non-random variation in diversification rates (Figure 2). Accordingly, the power of the whole-tree statistics to detect diversification rate variation should be enhanced by scaling the weight of nodal probabilities according to the size (i.e., species diversity) of their respective nodes. Because diversification is an exponential process, the contribution of each nodal probability is scaled by the natural logarithm of its diversity. The cumulative whole-tree probability based on the product of weighted ERM nodal probabilities,  $M_{\Pi}^*$ , involves first calculating the product of weighted ERM nodal probabilities,  $\Pi^*$ , for the observed tree and the set of simulated trees using the equation

$$(2) \quad \Pi^* = \frac{\prod_{i=1}^{n-1} \ln(n_i) \ln(P_i)}{\sum_{i=1}^{n-1} \ln(n_i)}.$$

(Recall that the sum of the natural logarithms of the ERM nodal probabilities is equivalent to taking their product.) The cumulative whole-tree probability,  $M_{\Pi}^*$ , is simply the frequency of simulated trees with  $\Pi^*$  values less than that of the observed tree. Similarly, the cumulative whole-tree probability based on the sum of weighted ERM nodal probabilities,  $M_{\Sigma}^*$ , involves calculating the sum of weighted ERM nodal probabilities,  $\Sigma^*$ , using the equation

$$(3) \quad \Sigma^* = \frac{\sum_{i=1}^{n-1} \ln(n_i) P_i}{\sum_{i=1}^{n-1} \ln(n_i)},$$

where  $n_i$  is the diversity of internal node  $i$ , and  $P_i$  is its corresponding ERM nodal probability derived using equation (1). Given two trees with the same number of tips but different topological shapes, the more asymmetric tree will contain a greater proportion of nodes that are relatively large compared with the more balanced tree. Accordingly, the denominators in equations (2) and (3) normalize the summation of  $\ln(n_i)$  over different tree shapes.



*Figure 2.* The ability to detect non-ERM diversification increases with tree size. The plots were generated by initiating a stochastic ERM-branching process from a single species with the diversification-rate parameter,  $\lambda$ , set initially to 1. After the first branching event, a diversification-rate shift of magnitude  $A$ , where  $A \in \{1, 2, 3, 4\}$ , was applied deterministically to one of the two lineages descended from the root node. The process was terminated when trees reached size  $N$ , where  $N \in \{1, 2, 3, \dots, 10, 15, 20, \dots, 40, 60, 80, 100\}$ . Each combination of parameter settings (magnitude of diversification-rate difference, tree size) was replicated 100 000 times, and, for each tree generated from each such replicate, the ERM nodal probability was calculated for the root node using equation (1). The graphs plot the mean, standard deviation, and various percentiles of the ERM nodal probabilities (where a percentile is the ERM  $P$ -value corresponding to the simulated tree for which  $x\%$  of the set of simulated trees had lower  $P$ -values, where  $x \in \{50, 80, 90, 95\}$ ). The plots within each of the four graphs (corresponding to a set of simulations under a given value of  $A$ ) are concave, with the ERM nodal probabilities for the root node decreasing in value with increasing tree size. For a given value of  $A$ , ERM  $P$ -values are clustered more tightly around small values for larger trees. Under a diversification-rate difference of three, for example, we are much more likely to obtain a  $P$ -value of  $< 0.1$  for  $N = 100$  than for  $N = 10$ , indicating that larger nodes provide more reliable evidence of non-ERM diversification. Note that the apparently stochastic wobbling of the percentile plots near the  $y$ -axis is actually a manifestation of the ‘discreteness’ problem. For a tree of a given size, only a finite number of discretely valued diversity partitions can be realized; accordingly, only a finite number of  $P$ -values can be realized by their corresponding nodal probabilities. As expected, the discreteness problem is most pronounced for trees of small size.

### 3.2 The relative sensitivity to diversification rate variation at different phylogenetic scales

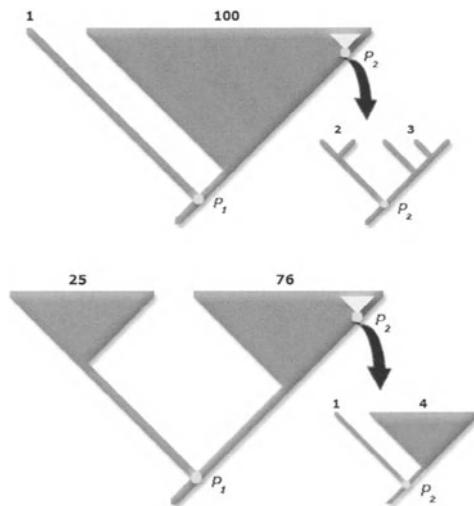
Our motivation for developing the whole-tree methods described above is to increase the statistical power of tests to detect diversification rate variation. Power is the ability of a test to reject a null hypothesis when it is false. Nodal

ERM probabilities are the most appropriate measure for tests of differential diversification at individual nodes. Accordingly, we expect the combination of these values — as implemented by the  $M$  statistics — to provide tests of the ERM model that are exceptionally sensitive to diversification rate variation within whole (super)trees.

Of course, the power of a test is contingent on the nature of the particular alternative hypothesis under consideration. Because there are innumerable possible alternatives to equiprobable diversification rates (frequent rate shifts dispersed throughout the tree or infrequent rate shifts occurring near the base of the tree, among others), it is unrealistic to expect any single statistic to be maximally powerful in all scenarios involving differential diversification. Given the multitude of possible and biologically relevant alternatives to ERM cladogenesis, several different statistics are required. The  $M$  statistics are intended to provide differential sensitivity to asymmetry arising at different phylogenetic scales (i.e., the relative nodal depth in the tree), permitting their application to a corresponding range of associated evolutionary processes.

The manner in which each statistic summarizes information from individual nodes (i.e., ERM probabilities) will determine the type of diversification rate variation (i.e., the alternative hypothesis) to which it is most sensitive. By considering how the different  $M$  statistics differentially summarize ERM nodal probabilities, we can theoretically characterize their differential sensitivity to different patterns of diversification rate variation without performing the simulations necessary for a complete characterization of their relative power.

Although  $M_N$  and  $M_\Sigma$  both consider the relative asymmetry of all internal nodes, these statistics nevertheless exhibit differential sensitivity to large-scale asymmetry. To understand the source of this difference, recall that the potential magnitude of diversity partitions is greater at more inclusive nodes. Consider, for example, that the most extreme diversity partition of an  $N$ -species tree is a split of  $1:(N - 1)$ , which can only be realized at the root; the next most extreme partition,  $2:(N - 2)$ , can only be realized at the root or at the node just above the root, and so on. Accordingly, the most extreme nodal probabilities (i.e., the smallest) can only be generated by large-scale asymmetry. These extreme probabilities will have a relatively large effect on  $M_N$  because calculation of the statistic involves their multiplication. By contrast,  $M_\Sigma$  combines nodal probabilities additively, such that the impact of such extreme probabilities is greatly diminished, allowing nodal probabilities associated with small-scale asymmetry to make a more equitable contribution to the whole-tree probability under this statistic.



**Figure 3.** The differential sensitivity of the whole-tree tests to diversification-rate variation manifested at different phylogenetic scales. Trees *A* and *B* exhibit substantial differences in large-scale phylogenetic asymmetry: *A* has a basal split of 1:100 ( $P_1 = 0.02$ ,  $\ln P_1 = -3.91$ ) versus a 25:76 split in *B* ( $P_1 = 0.5$ ,  $\ln P_1 = -0.69$ ). Now, imagine that the only other difference in asymmetry between the two trees is restricted to a five-species subtree that has a 2:3 split in *A* ( $P_2 = 1.0$ ,  $\ln P_2 = 0$ ) and a 1:4 split in *B* ( $P_2 = 0.5$ ,  $\ln P_2 = -0.69$ ). The number of such asymmetric five-species subtrees that would be required by each whole-tree statistic to identify *B* as more asymmetric than *A* can be used to characterize their relative sensitivity to small-scale phylogenetic asymmetry.  $M_\Sigma$  identifies *B* as more asymmetric with just a single asymmetric five-species subtree ( $P_{1A} + P_{2A} = 1.02$ ;  $P_{1B} + P_{2B} = 1.0$ );  $M_\Sigma^*$  requires three or more equivalent differences;  $M_\Pi$  requires five or more equivalent differences ( $\ln P_{1A} - \ln P_{1B} = -3.22$ ,  $\ln P_{2A} - \ln P_{2B} = 0.69$ ); and  $M_\Pi^*$  requires 14 or more equivalent differences. For comparison,  $I_C$  requires 25 or more equivalent differences in small-scale asymmetry, whereas  $B_I$  identifies *B* as far more asymmetric than *A* with only a single such difference. Thus, the sensitivity of the whole-tree statistics to diversification-rate variation occurring at large phylogenetic scales is approximately  $B_I < M_\Sigma < M_\Sigma^* < M_\Pi < M_\Pi^* < I_C$ .

Predictably, the behavior of the weighted whole-tree statistics,  $M_\Pi^*$  and  $M_\Sigma^*$ , is similar to that of their equally weighted counterparts. However, because the contribution of each ERM nodal probability to these whole-tree statistics is weighted by the size of its corresponding node, and because larger nodes are realized deeper in the tree,  $M_\Pi^*$  and  $M_\Sigma^*$  are more sensitive to diversification rate variation at larger phylogenetic scales. Accordingly, the relative sensitivity of the  $M$  statistics to large-scale diversification rate variation can be approximately characterized as  $M_\Sigma < M_\Sigma^* < M_\Pi < M_\Pi^*$  (Figure 3).

### 3.3 Assessing the statistical behavior of the whole-tree statistics

We performed a simulation study to characterize the relative power of the five whole-tree statistics ( $M_{\Sigma}$ ,  $M_{\Sigma}^*$ ,  $M_{II}$ ,  $M_{II}^*$ , and  $M_R$ ) and two previously proposed balance indices:  $I_C$  (Colless, 1982; Heard, 1992) and  $B_I$  (Shao and Sokal, 1990). Our decision to compare the  $M$  statistics with these two balance metrics is based on several considerations.  $I_C$  is both the most commonly used index (e.g., Mooers and Heard, 1997) and is also very well characterized mathematically (e.g., Heard, 1992; Rogers 1993, 1994, 1996). By contrast, our inclusion of  $B_I$  is motivated by the finding that it is the most powerful of the balance indices (Kirkpatrick and Slatkin, 1993; but see Agapow and Purvis, 2002).

The ability of the seven statistics to detect diversification rate variation was assessed by a simulation design that involved growing trees under a variety of non-ERM conditions intended to simulate plausible and potentially biologically interesting models of cladogenesis. In general, trees were grown under a continuous-time, discrete-state, stochastic branching process in which splitting events were assumed to be both instantaneous and dichotomous. The probability of a branching event was assumed to be independent between tips in a growing tree, with rate shifts being equally likely to involve an increase or a decrease in diversification rate. If no rate shift occurred, a given tip retained the diversification rate of its ancestor.

Diversification rate shifts were applied under three general models of cladogenesis. Under the gradualist model, rate shifts could occur at any instant in time and were inherited by both daughter species. Alternatively, two different punctuated models constrained rate shifts to occur at speciation events, with either one or both daughter species having a chance of experiencing a rate shift. For each evolutionary model, we explored the effects of varying the frequency and magnitude of rate shifts in trees of various sizes. Average diversification rate shift values,  $\lambda$ , included two-, four-, eight-, and 16-fold increases in diversification rate, which were applied under a range of frequencies (0.01, 0.1, 0.2, 0.3, 0.4, 0.5). The branching process was terminated when trees reached the desired size,  $N$ , where  $N \in \{10, 15, 20, 25, 30, 35, 40, 60, 80, 100\}$ . Every permutation of the set of simulation parameters (evolutionary model, rate distribution, tree size, and frequency and magnitude of rate shifts) was replicated 100 000 times, calculating the value for each of seven statistics for each tree generated from each replicate. Power was calculated as the proportion of the replicates in which the null hypothesis of no among-lineage diversification rate variation was correctly rejected at the conventional  $\alpha = 0.05$ .

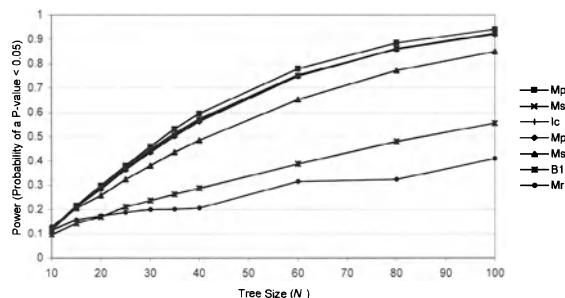


Figure 4. The effect of tree size on the power of several whole-tree methods to detect diversification-rate variation.

Several results of the simulation study were as predicted. First, the use of Monte Carlo simulation to assess significance of the various whole-tree statistics ensured appropriate Type I error rates. The plots for each statistic intersected the  $y$ -axes at  $P \approx 0.05$  (the nominal level of  $\alpha$ ) when the average diversification rate shift,  $\lambda$ , was 1 (i.e., when the null hypothesis was true). Second, the power of the whole-tree statistics to detect diversification rate variation consistently scaled with tree size (Figure 4). This result is consistent both with theoretical expectations (Figure 2) and findings of previous simulation studies (e.g., Kirkpatrick and Slatkin, 1993; Kubo and Iwasa, 1995; Paradis, 1997, 1998a, b; Agapow and Purvis, 2002) and emphasizes the potential of typically large supertrees to facilitate the study of diversification rate variation. Finally, the observed behavior of the various whole-tree statistics under various rate-shift parameterizations was also unsurprising: power predictably scaled with increases in both the frequency and magnitude of rate shifts applied.

Somewhat more surprising was the response of some whole-tree statistics to various combinations of frequency and magnitude of diversification rate shifts. For instance, we might expect that simulations involving large shifts occurring at low frequencies would enhance the relative power of the  $M_R$  statistic given its inherent sensitivity to large-scale diversification rate variation. Similarly, we might predict that the relative performance of the  $M_\Sigma$  or  $B_1$  statistics would be enhanced under conditions involving shifts of small magnitude occurring at relatively high frequencies. Curiously, and despite their rather compelling theoretical basis, no unambiguous patterns supporting these behaviors emerged from the simulation study. A thorough consideration of such intriguing anomalies is beyond the scope of the present analysis but will be treated elsewhere (Moore and Chan, in prep.).

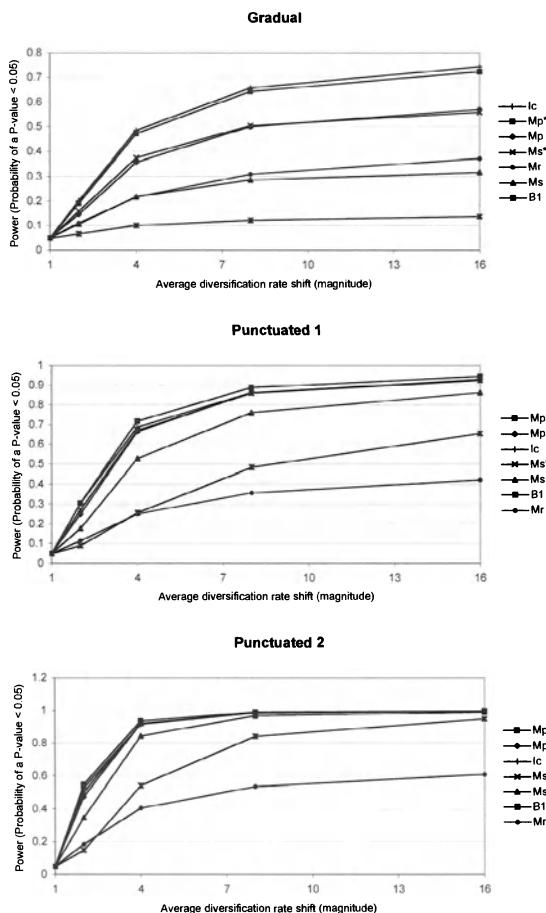


Figure 5. The effects of evolutionary model on the power of several whole-tree methods to detect diversification-rate variation.

Perhaps one of the more interesting findings to emerge from the simulation study was the pronounced effect of the model of diversification on the relative power of the whole-tree methods to detect diversification rate variation. Figure 5 depicts the results for 100-species trees grown under one of three diversification models with rate shifts of various magnitude applied with a constant frequency of 0.1, which were sampled from a uniform rate-shift distribution. Overall, the power of all the statistics tended to be greatest under the Punctuated 2 Model, in which rate shifts were constrained to occur at speciation events with any change in rate inherited by both daughter species (Figure 5, lower graph). By contrast, power was noticeably lower

under the Gradual Model, in which rate shifts were free to occur at any time with any change in rate shared between both daughter species (Figure 5, upper graph). Finally, power was intermediate under the Punctuated 1 Model, in which only one of the descendant species inherited any change in rate (Figure 5, middle graph). Nevertheless, the absolute power of the methods was fairly high even under conditions least favorable to the whole-tree statistics (e.g., two of the statistics,  $M_{II}^*$  and  $I_C$ , detected a four-fold variation in diversification rates correctly about 50% of the time under the Gradual Model at a very low diversification rate shift frequency).

The conditional nature of conclusions regarding the statistical power of these tests should be emphasized. Despite this caveat, several generalities held over a wide range of the considerable parameter space we explored. Apart from a limited number of extreme conditions, the performance of the  $M_R$  and  $B_I$  statistics was uniformly poor. Given its widely accepted status as the most powerful statistic (based on the particular conditions simulated by Kirkpatrick and Slatkin, 1993), the poor performance of  $B_I$  was somewhat surprising. By contrast, the  $M_{II}^*$  statistic consistently exhibited maximal (or nearly maximal) power under the vast majority of the simulations.

### 3.4 Detecting diversification rate variation in primates

The whole-tree  $M$  statistics described above were used to assess diversification rate variation in a published supertree of primates (Purvis, 1995). Because these data were analyzed for illustrative purposes only, no attempt was made to account for the effect of phylogenetic uncertainty on the results (e.g., Donoghue and Ackerly, 1996; Huelsenbeck *et al.*, 2000b). In addition to analyzing the complete primate tree, we also performed analyses on several clades of primates to facilitate comparison both with the findings of previous temporal studies of diversification rate variation in this group (Purvis *et al.*, 1995) and also with results presented in Section 4.3. Note that inference of diversification rate variation in these clades is somewhat confounded: shifts within more nested clades will influence estimates obtained for more inclusive clades. Accordingly, these results should be interpreted cautiously (Purvis *et al.*, 1995). Results derived with the whole-tree methods were compared again with those of the tree-shape indices  $I_C$  and  $B_I$ . All analyses were performed with SYMMETREE, with relevant details and results summarized in Table 1.

Three general findings merit comment. First, the primate tree contains 203 species and is ~80% resolved, illustrating the ability of the whole-tree methods (and their implementation in SYMMETREE) to contend with moderately large and incompletely resolved trees. Second, analysis of the entire primate clade failed to detect significant among-lineage diversification

*Table 1.* Probability values corresponding to tests of ERM cladogenesis in various primate clades as derived by Monte Carlo simulation of the null distribution for each statistic. All results were obtained using the SYMMETREE program. The null distribution for each statistic was generated with a sample of 100 000 ERM topologies for each tree size. Uncertainty associated with polytomies was assessed by generating 100 000 random resolutions under the size-sensitive ERM taxon-addition algorithm, providing the upper and lower bounds of the confidence interval. These bounds, the “high” and “low” values (for high and low asymmetry), correspond to the tail probabilities for the .025 and .975 frequentiles, respectively. Note that the sensitivity of the whole-tree statistics to large-scale diversification rate variation increases to the right across a given row (i.e.,  $B_1 < M_\Sigma < M_\Sigma^* < M_{\Pi} < M_{\Pi}^* < I_C$ ). Percent resolution was calculated as  $k / (N - 1)$ , where  $k$  is the number of nodes in a tree of  $N$  species; this value assumes implicitly that the underlying phylogeny is strictly dichotomous (i.e., that all polytomies are “soft”; *sensu* Maddison, 1989).

| taxon                | tree<br>size | resolution | $B_1$              | $M_\Sigma$         | $M_\Sigma^*$       | $M_{\Pi}$          | $M_{\Pi}^*$        | $I_C$              |
|----------------------|--------------|------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|
|                      |              |            | high<br>low        | high<br>low        | high<br>low        | high<br>low        | high<br>low        | high<br>low        |
| all primates         | 203          | 79         | 0.00020<br>0.09481 | 0.00414<br>0.18733 | 0.04097<br>0.30751 | 0.00468<br>0.12587 | 0.04004<br>0.20208 | 0.21772<br>0.32138 |
| hominoids            | 14           | 85         | 0.00074<br>0.01481 | 0.00507<br>0.03833 | 0.01956<br>0.08997 | 0.01063<br>0.07831 | 0.04627<br>0.17608 | 0.05123<br>0.18630 |
| strepsirrhines       | 39           | 82         | 0.14391<br>0.77485 | 0.12380<br>0.67503 | 0.12380<br>0.33586 | 0.24500<br>0.71871 | 0.48850<br>0.82819 | 0.67694<br>0.89569 |
| New World<br>monkeys | 65           | 72         | 0.01541<br>0.73206 | 0.32875<br>0.97304 | 0.65913<br>0.98205 | 0.44013<br>0.96326 | 0.65815<br>0.96834 | 0.78990<br>0.95474 |
| Old World<br>monkeys | 80           | 81         | 0.00168<br>0.18815 | 0.00134<br>0.09342 | 0.00473<br>0.08558 | 0.00045<br>0.02488 | 0.00384<br>0.03825 | 0.13364<br>0.30341 |

rate variation. However, significant diversification rate variation was detected in separate analyses of both hominoids and Old World monkeys. These findings are largely consistent with those reported by Purvis *et al.* (1995), who detected diversification rate variation within both of these clades using temporal methods. Finally, close inspection of the  $P$ -values for the various whole-tree statistics supports their predicted behavior with respect to diversification rate variation manifest at different phylogenetic scales. The statistics in Table 1 are arranged by their predicted sensitivity to large-scale diversification rate variation (i.e., in the order  $B_1 < M_\Sigma < M_\Sigma^* < M_{\Pi} < M_{\Pi}^* < I_C$ ). Looking across a row for any group reveals a trend in the  $P$ -values; for example, the probabilities for Old World monkeys tend to decrease from  $B_1$  to  $M_{\Pi}$  and then increase from  $M_{\Pi}^*$  to  $I_C$  (with some shuffling of the order of the statistics resulting from differences in the their absolute power under the particular manner in which the null hypothesis was violated in these data). The most extreme  $P$ -value (i.e., the smallest) obtained for this clade was returned by  $M_{\Pi}$ , suggesting that diversification

rate variation in the Old World monkey tree likely occurred at an intermediate phylogenetic scale.

#### 4. Locating shifts in diversification rate

Having provided a means with which to answer the question, “*Has* a given tree experienced significant diversification rate variation among its branches?” in the preceding section, we now address its inevitable sequel: “*Where* have significant shifts in diversification rate occurred in this tree?” Despite its obvious biological significance, this problem has received remarkably little attention (however, Nee *et al.* (1992, 1994b, 1996) developed an approach incorporating temporal information that has been applied to this problem, which we consider in some detail below). By contrast, considerable attention has focused on methods to test hypotheses that specify the location and direction of diversification rate shifts (i.e., “key-innovation” hypotheses<sup>3</sup>). Fortunately, several developments in this hypothesis-testing realm are directly relevant to the issue of localizing shifts in diversification rate. Of particular importance is the iterative maximum likelihood model-fitting approach proposed by Sanderson and Donoghue (1994; see also Sanderson and Bharathan, 1993; Sanderson, 1994; Sanderson and Wojciechowski, 1996).

Following Sanderson and Donoghue (1994), our approach to detecting shifts in diversification rate is developed in a likelihood framework that evaluates the relative fit of models with one or more rate parameters distributed over different parts of a three-taxon tree and assumes an underlying ERM (Yule) branching process. However, our implementation is both significantly simplified (we evaluate only one- and two-rate parameter models and do not integrate their likelihood over all internal branching times) and also substantially generalized (we iterate three-taxon evaluations over all internal branches to survey the whole tree for diversification rate shifts).

In outline, the basic goal is to assess the probability of a shift along the lone internal branch of a given three-taxon tree comprising an outgroup clade and the two basal-most subclades of the ingroup clade. The probability of a diversification rate shift along the internal branch is returned by a shift

<sup>3</sup> Although related, these inference problems are nevertheless distinct. The evaluation of key innovations entails a hypothesis-testing framework in which the location and direction of a diversification rate shift is specified by the hypothesis under consideration (without any knowledge that the tree exhibits significant among-lineage diversification rate variation). By contrast, the search for significant shifts in diversification rate entails a data-exploration framework in which only the existence of significant among-lineage diversification rate variation is specified (without any knowledge of the location or direction of the associated rate shifts).

statistic, which is calculated as a function of two likelihood ratios. One likelihood ratio is calculated at the root of the three-taxon tree (involving the diversity partition between the outgroup and ingroup clades), the other at the root of the ingroup clade (involving the diversity partition between the left and right ingroup clades). Each likelihood ratio compares the likelihood of realizing the observed diversity partition between the two sister clades under a homogeneous (one-rate parameter) model (in which both groups have the same branching rate) versus that under a heterogeneous (two-rate parameter) model (in which the two groups have different branching rates). Different shift statistics can be developed by variously combining information from the resulting inclusive and nested likelihood ratios. Before explicitly deriving these shift statistics, we first review both the details of calculating the likelihoods under one- and two-rate parameter models and also the means of assessing their relative fit to the data using the likelihood ratio.

If the ERM branching process is initiated with a single species and allowed to run for a period of time,  $t$ , with a branching probability,  $\lambda$ , the likelihood of realizing  $N$  species is (Harris, 1964)

$$(4) \quad P(N | \lambda, t) = e^{-\lambda t} (1 - e^{-\lambda t})^{N-1}.$$

Accordingly, the likelihood of realizing  $N$  species partitioned between the left and right descendants of a single node (with  $\ell$  and  $r$  species, respectively) under a uniform branching probability after time,  $t$ , is

$$(5a) \quad P(\ell, r | H_O) = \frac{P(\ell | \lambda, t)P(r | \lambda, t)}{\sum_{i=1}^{N-1} P(i | \lambda, t)P(N-i | \lambda, t)}.$$

The Markov property of the ERM branching process allows the probabilities for different parts of the tree (such as the two terms in the numerator) to be multiplied. Substituting the expression from equation (4) with  $t = 1$  gives the following expansion

$$(5b) \quad P(\ell, r | H_O) = \frac{\left(e^{-\lambda}(1-e^{-\lambda})^{\ell-1}\right)\left(e^{-\lambda}(1-e^{-\lambda})^{r-1}\right)}{\sum_{i=1}^{N-1} \left(e^{-\lambda}(1-e^{-\lambda})^{i-1}\right)\left(e^{-\lambda}(1-e^{-\lambda})^{N-i-1}\right)}.$$

This equation provides the likelihood of observing a partition of  $\ell$  and  $r$  species (where  $\ell + r = N$ ) under  $H_O$ , the homogeneous, one-rate parameter

model. Similarly, the likelihood of observing a partition of  $\ell$  and  $r$  species under the heterogeneous, two-rate parameter model,  $H_A$ , is

$$(6a) \quad P(\ell, r | H_A) = \frac{P(\ell | \lambda_\ell, t)P(r | \lambda_r, t)}{\sum_{i=1}^{N-1} P(i | \lambda_\ell, t)P(N-i | \lambda_r, t)}.$$

Again, substituting the expression from equation (4) with  $t = 1$  gives the expansion

$$(6b) \quad P(\ell, r | H_A) = \frac{\left(e^{-\lambda_\ell} (1 - e^{-\lambda_\ell})^{\ell-1}\right)\left(e^{-\lambda_r} (1 - e^{-\lambda_r})^{r-1}\right)}{\sum_{i=1}^{N-1} \left(e^{-\lambda_\ell} (1 - e^{-\lambda_\ell})^{i-1}\right)\left(e^{-\lambda_r} (1 - e^{-\lambda_r})^{N-i-1}\right)},$$

The denominators in equations (5) and (6) normalize their respective probabilities by defining the relevant probability space. Specifically, this pertains to the sum of the products for all possible partitions of  $N$  into  $\ell$  and  $r$  species.

The relative fit of the one- and two-rate parameter models to the observed diversity partition is assessed by the difference in the natural logarithm of their respective likelihood values: the log-likelihood ratio (hereafter, simply “likelihood ratio”) of the homogeneous and heterogeneous diversification rate models,  $LR_{H_A:H_0}$ , is, therefore, calculated as

$$(7) \quad LR_{H_A:H_0} = \ln \left( \frac{\sum_{i=1}^{N-1} P(n_i | \lambda_\ell, t)P(N-n_i | \lambda_r, t)}{\sum_{i=1}^{N-1} P(n_i | \lambda_\ell, t)P(N-n_i | \lambda_r, t)} \right) - \ln \left( \frac{\sum_{i=1}^{N-1} P(n_i | \lambda, t)P(N-n_i | \lambda, t)}{\sum_{i=1}^{N-1} P(n_i | \lambda, t)P(N-n_i | \lambda, t)} \right).$$

As the value of likelihood ratio increases, the evidence increasingly favors acceptance of the heterogeneous model in which the left and right descendants of the node in question diversified under two distinctly different rates,  $\lambda_\ell$  and  $\lambda_r$ , respectively.

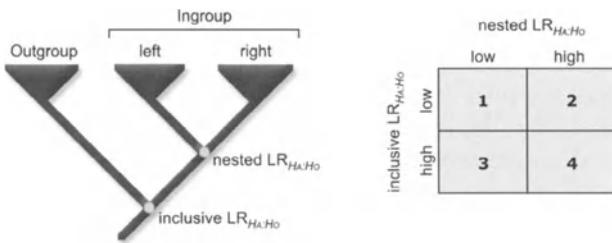
#### 4.1 Presentation of the shift statistics

Having detailed the calculation of likelihoods under the homogeneous and heterogeneous models (and their relative fit with the likelihood ratio), we now have the necessary tools to construct tests to locate significant shifts in

diversification rate. Consider a pair of sister taxa,  $L$  and  $R$ , with  $\ell$  and  $r$  species, respectively (where  $\ell < r$ ). After calculating the likelihood of realizing a partition of  $\ell$  and  $r$  species under both the homogeneous and heterogeneous models, we then calculate the difference in their log likelihoods (i.e., the likelihood ratio of  $H_A:H_O$ ). The discovery of a large likelihood ratio would provide evidence that  $L$  and  $R$  diversified under two distinctly different rates,  $\lambda_L$  and  $\lambda_R$ , respectively. We might interpret this as evidence of an increase in diversification rate along the internal branch leading to  $R$  (i.e., the *stem* branch subtending the  $R$  clade; *sensu* Doyle and Donoghue, 1993; Magallón and Sanderson, 2001). However, this interpretation relies on several assumptions, including the key assumption that the diversity of the more diverse group,  $R$ , was achieved stochastically under a constant rate,  $\lambda_R$  (e.g., Rakiow, 1986; Sanderson and Donoghue, 1996)<sup>4</sup>. It is possible that an apparent shift in rate along the branch leading to  $R$  could be an artifact of a rate shift that occurred within  $R$ . This “trickle-down” problem occurs because a bona fide increase in diversification rate along a given internal branch will exert an influence on diversity comparisons made at more inclusive nodes. Accordingly, a local shift in rate is effectively conducted down the tree, creating the illusion of local rate shifts at neighboring internal branches (see Figure 6).

To discriminate between such illusory and real rate shifts, therefore, we must expand the scope of our evaluation to incorporate information not only from the node subtended by  $L$  and  $R$  but also from the root node of  $R$ . Evaluation of these two hierarchically nested nodes thus entails a three-taxon framework comprising an outgroup clade and the two basal subclades that

<sup>4</sup> The other key assumption concerns the inferred direction of the shift in rate: as two-taxon statements, sister-group comparisons are inherently non-directional (e.g., Jensen, 1990; Doyle and Donoghue, 1993; Sanderson and Bharathan, 1993; Sanderson and Donoghue, 1994, 1996; Sanderson and Wojciechowski, 1996). In other words, the observation that clade  $R$  contains significantly more species than its sister group,  $L$ , can be explained by postulating either a rate increase in  $R$  and/or a rate decrease in  $L$ . In principle, increases and decreases in diversification rate are likely to have occurred with equal frequency throughout evolutionary history. Nevertheless, our method ignores shifts associated with significant decreases in diversification rate because the detection of such events on the basis of extant diversity is highly problematic given the associated loss of relevant phylogenetic information. That is, while we do not deny the existence of significant decreases in diversification rate, we are unlikely to detect these events because their occurrence effectively ensures the erasure of the evolutionary history necessary for their discovery. The probability that an entire clade will go extinct is governed by the relative extinction rate,  $\epsilon$ , which is simply the extinction rate divided by the speciation rate (e.g., Kendal, 1948; Harris, 1964; Nee *et al.*, 1994b; Magallón and Sanderson, 2001). As  $\epsilon$  increases, it becomes increasingly likely that a clade will perish before the present; when  $\epsilon \geq 1$ , the probability of complete extinction is one. Evidence from the fossil record suggests that  $\epsilon$  has historically been quite high for most groups (e.g., Stanley, 1979; Hulbert, 1993). Recall that a significant decrease in the net diversification rate,  $\lambda$ , entails a significant decrease in speciation rate and/or a significant increase in extinction rate. Such a decrease in  $\lambda$  will therefore cause a corresponding increase in  $\epsilon$ , which will greatly increase the probability that the clade will go extinct before the present. Accordingly, if a significant decrease in rate actually occurred in a given group, there would likely be no record of such an event in the relationships among extant species.



**Figure 6.** Locating significant shifts in diversification rate in the context of a three-taxon tree. Note that the tree has been rendered in left-light rooting order (Furnas, 1984), such that the more diverse clade is swiveled to the right of every node. Detection of a rate shift along the internal or target branch entails calculation and evaluation of likelihood ratios under the homogeneous and heterogeneous models,  $LR_{H_A:H_O}$ , at both the inclusive and nested nodes. The fit of the heterogeneous model to an observed diversity partition at a given node increases with the value of the likelihood ratio. Inspection of the inclusive and nested likelihood ratios entails one of four possible interpretations. Scenarios 1 and 2 indicate that no rate shift occurred along the target branch (although scenario 2 is consistent with a rate shift within the ingroup, which will be assessed as the three-taxon evaluation is iterated up the tree). By contrast, the large likelihood ratios at the inclusive nodes in scenarios 3 and 4 suggest that a rate shift might have occurred along the target branch. In scenario 4, however, the large value of the nested likelihood ratio suggests that rates within the ingroup are significantly heterogeneous. Accordingly, the apparent rate shift along the target branch is likely an artifact of a subsequent rate shift within the ingroup. Thus, scenario 3 represents a bona fide rate shift along the target branch, whereas scenario 4 illustrates the “trickle-down” problem.

together form the ingroup. In outline, the likelihood of a shift along the internal branch of the three-taxon tree (which is based on the likelihood ratio for the observed diversity partition between the outgroup and ingroup clades) must be conditioned by the likelihood of a rate shift within the ingroup (which is based on the likelihood ratio for the observed diversity partition between the left and right ingroup clades). There are many ways one might conceive of conditioning the inclusive likelihood ratio by the nested likelihood ratio, each variant corresponding to a different likelihood ratio-based shift statistic. Indeed, many shift statistics could be imagined that are based on expressions of the data other than their likelihood ratio. In fact, we have developed and experimented with several such alternative shift statistics (see below). Nevertheless, we focus on two shift statistics based on nested likelihood ratios because of their advantageous statistical properties.

The first shift statistic,  $\Delta_1$ , simply takes the difference in likelihood ratios under the homogeneous and heterogeneous models assessed at the inclusive and nested nodes. It is calculated as

$$(8) \quad \Delta_1 = \left( LR_{H_A:H_O} n_{OG} : n_{IG_L} \right) - \left( LR_{H_A:H_O} n_{IG_L} : n_{IG_R} \right),$$

where  $n_i$  is the number of species in group  $i$ , and  $LR_{H_A:H_O:n_i:n_j}$  is the likelihood ratio of observing a diversity partition  $n_i:n_j$  under the homogeneous and heterogeneous models derived using equation (7). The idea is to condition the evidence for a shift at the inclusive node (as reflected by the likelihood ratio of the observed diversity partition between the ingroup and outgroup clades,  $n_{OG}:n_{IG}$ ) by the evidence of a shift at the nested node (as reflected by the likelihood ratio of the observed diversity partition between the left and right ingroup clades,  $n_{IG_L}:n_{IG_R}$ ), thereby reducing the probability of erroneously attributing a local rate shift to the internal branch because of a rate shift within the ingroup clade.

The second shift statistic,  $\Delta_2$ , is more complicated. Rather than conditioning the inclusive likelihood ratio on the nested likelihood ratio, it attempts to adjust the ingroup diversity used in calculating the inclusive likelihood ratio. The adjusted ingroup diversity excludes the number of ingroup species that can be attributed to a rate increase along the internal branch. This value is calculated as the total ingroup diversity minus the product of the probability of a rate shift at the internal branch, multiplied by the number of species attributable to that shift. The  $\Delta_2$  shift statistic is expressed as

$$(9a) \quad \Delta_2 = \left( LR_{H_A:H_O} n_{OG} : n_{IG^*} \right),$$

where

$$(9b) \quad IG^* = n_{IG} - \left( \frac{\left( LR_{H_A:H_O} n_{IG_L} : n_{IG_R} \right)}{\left( LR_{H_A:H_O} n_{IG_L} : n_{IG_R} \right) + 1} \right) \left( n_{IG} - \max(n_{OG}, 2n_{IG_L}) \right).$$

The second term in equation (9b) constrains the adjusted ingroup diversity to assume the larger of two values: the outgroup diversity or two times the diversity of the less diverse (left) ingroup clade. This constraint is imposed to avoid overcorrecting the ingroup diversity in cases for which there is little evidence of a shift along the internal branch. The ERM  $P$ -values associated with the shift statistics  $\Delta_1$  and  $\Delta_2$  are assessed by numerical analysis: the cumulative probability of obtaining a shift statistic value as or more extreme than that derived for the observed tree (using equation (8) or (9)) is calculated using the statistic value for the observed topology and the known probabilities of different topologies under the ERM model.

## 4.2 Assessing the statistical behavior of the shift statistics

We performed a simulation study to explore the behavior of several shift statistics using a simple experimental design in which a rate shift was applied to either the inclusive and/or nested node of a three-taxon tree. The power and bias of the various shift statistics were assessed by their respective abilities to correctly or incorrectly reject the null hypothesis that no rate shift occurred along the internal (target) branch. Specifically, trees were generated under an ERM branching process initiated from a single species with the branching rate parameter,  $\lambda$ , set to 1. During the growth of a simulated tree, a diversification rate shift of a specified magnitude  $A$ , where  $A \in \{2, 4, 6\}$ , occurred deterministically under three different treatments: 1) a shift was applied to the inclusive node (i.e., occurring immediately after the first branching event); 2) a shift was applied to the nested node (i.e., occurring immediately after the first branching event within the ingroup); or 3) a shift was applied both to the inclusive and nested nodes (i.e., occurring immediately after the first and second branching events). The process was terminated when the trees reached the desired size,  $N$ , where  $N \in \{100, 200, 400\}$ . Every permutation of the set of simulation parameters (tree size, magnitude of rate shifts, and location of rate shifts) was replicated 10 000 times, calculating the value for each of seven shift statistics for each tree generated from each replicate. Power and Type I error rates were calculated as the proportion of the replicates in which the null hypothesis of no rate shift along the target branch was correctly or incorrectly rejected, respectively.

We compared the performance of our two likelihood ratio statistics,  $\Delta_1$  and  $\Delta_2$ , to one existing and four other new shift statistics:

1.  $NP$ , the ERM nodal probability proposed by Slowinski and Guyer (1989a, b) was calculated for the inclusive node using equation (1).
2.  $\Delta_N$ , calculated as the difference in “raw” diversity contrasts at the inclusive and nested nodes; that is,  $\Delta_N = ((n_{IG} - n_{OG}) - (n_{IGR} - n_{IG}))$ .
3.  $\Delta_R$ , calculated as the difference in diversification rate contrasts at the inclusive and nested nodes, where the maximum likelihood estimates of diversification rates are calculated as  $\hat{\lambda} = (\ln(n))^{-1}$  (Sanderson and Donoghue, 1996); accordingly,  $\Delta_R = ((\hat{\lambda}_{IG} - \hat{\lambda}_{OG}) - (\hat{\lambda}_{IGR} - \hat{\lambda}_{IG}))$ .
4.  $\Delta_{NP}$ , calculated as the difference in the two ERM nodal probabilities calculated at the inclusive and nested nodes, which is somewhat similar to the procedure outlined by Nee and Harvey (1994; see also Nee *et al.*, 1996; Mayhew, 2002).
5.  $\Delta_{1^*\omega}$ , calculated as for  $\Delta_1$ , but incorporates a scaling parameter,  $\omega$ , that weights the contribution of the nested likelihood ratio (the second term in

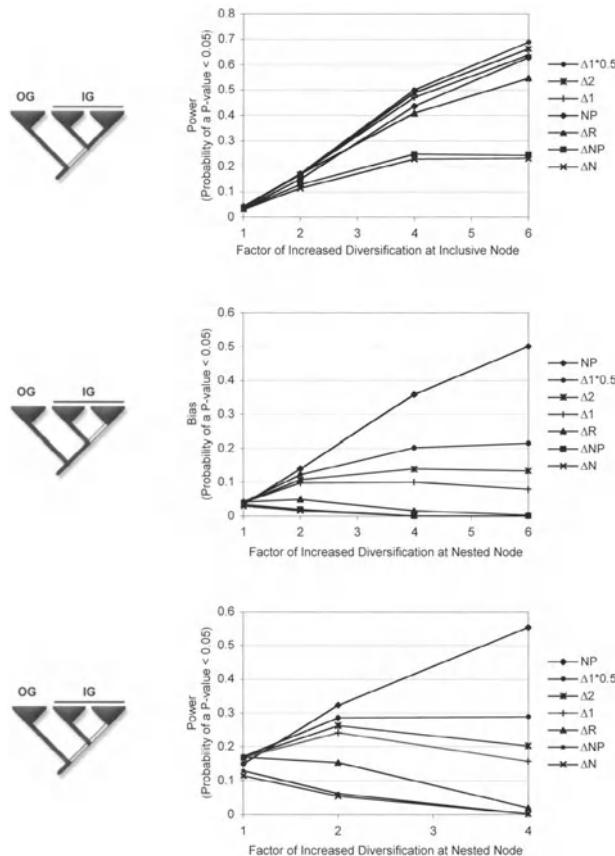


Figure 7. Results of a simulation study of the relative power and bias of several shift statistics in the three-taxon case. The locations of diversification rate shifts in the three-taxon trees are indicated as shaded branches (see text for details).

equation (8)) to the shift statistic; accordingly,  $\omega$  effectively indexes an infinite array of shift statistics, where  $\Delta_{1^*\omega}$  is identical to  $\Delta_1$  when  $\omega = 1$ ).

The upper graph in Figure 7 illustrates the ability of the shift statistics to detect diversification rate shifts of various magnitudes along the internal branch. These conditions correspond to the idealized case in which potentially confounding diversification rate shifts elsewhere in the tree have not occurred. The plots for each statistic intersect the y-axis at  $P \approx 0.05$  (the nominal level of  $\alpha$ ) where the diversification rate is raised by a factor of 1, indicating appropriate Type I error rates when the null hypothesis is true (as

expected under Monte Carlo simulation). For 100-species trees, the likelihood ratio-based shift statistics exhibit the greatest relative power, successfully detecting a four-fold rate increase in ~50% of the replicates, and a six-fold rate increase in ~65% of the replicates. The  $\Delta_{1*0.5}$  statistic slightly outperforms  $\Delta_2$ , which in turn slightly outperforms  $\Delta_1$ . All three likelihood ratio-based shift statistics enjoy an edge in power over Slowinski and Guyer's (1989a, b) *NP*, which is expected since these  $\Delta$  statistics possess greater resolution by virtue of incorporating more information. The other shift statistics,  $\Delta_R$ ,  $\Delta_{NP}$ , and  $\Delta_N$ , exhibit substantially lower power.

The middle graph in Figure 7 illustrates the bias of the various shift statistics associated with a diversification rate shift of various magnitudes within the ingroup (specifically, along the branch subtending the right ingroup clade). This simulation therefore assesses the relative sensitivity of the various shift statistics to the trickle-down problem. Because no rate increase occurs along the target branch, a completely unbiased statistic should exhibit a flat probability of rejecting the null hypothesis of ~0.05. As expected, *NP* is extremely biased, rejecting the null hypothesis almost as frequently as when a rate increase actually occurred at the target node (compare the plots for *NP* in the upper and middle graphs). The likelihood ratio-based shift statistics fare substantially better, exhibiting Type I error rates ranging between 10–20% under a four-fold rate increase within the ingroup and between 8–21% under a six-fold rate increase within the ingroup. Not surprisingly, the Type I error rates of the three likelihood ratio-based shift statistics mirror their relative power in the upper graph. Accordingly, the slight edge in power exhibited by  $\Delta_{1*0.5}$  translates into greater bias. The other two likelihood ratio-based shift statistics exhibited relatively low bias, with  $\Delta_1$  consistently outperforming  $\Delta_2$ . The remaining shift statistics,  $\Delta_R$ ,  $\Delta_{NP}$ , and  $\Delta_N$ , are substantially more conservative.

The lower graph in Figure 7 illustrates the ability of the shift statistics to detect a doubling in diversification rate along the internal (target) branch given a subsequent rate shift of varying magnitude within the ingroup clade. This simulation therefore assesses the power of the shift statistics in cases where the trickle-down problem applies. Because a doubling in rate is uniformly applied to the internal branch, the plot for each shift statistic intersects the *y*-axis at the ordinate value corresponding to its respective power under a two-fold rate increase in the upper graph. Note that, because a two-fold rate increase is consistently applied to the target node, a perfectly unbiased shift statistic would exhibit a flat power curve over the range of rate increases applied within the ingroup clade. Not surprisingly, *NP* exhibits the highest power under this scenario because shifts within the ingroup contribute to rejection of the null hypothesis; that is, it does a “good” job, albeit for the wrong reasons. The power plots for the likelihood ratio-based

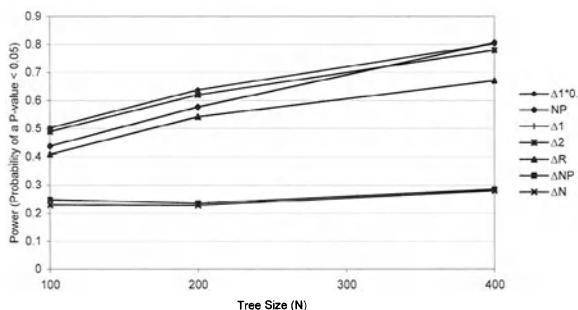


Figure 8. The effect of tree size on the power of several shift statistics to detect significant diversification rate shifts (see text for details).

shift statistics are substantially flatter, their rank order remaining unchanged:  $\Delta_1 < \Delta_2 < \Delta_{1 \cdot 0.5}$ . Because none of these  $\Delta$  shift statistics perfectly condition the inclusive likelihood ratio by the nested likelihood ratio, their power might be inflated slightly by a rate shift within the ingroup clade. Under these conditions, the  $\Delta_1$ ,  $\Delta_2$ , and  $\Delta_{1 \cdot 0.5}$  statistics appear to slightly undercondition the inclusive by the nested likelihood ratio. This behavior can be seen by comparing the plots of  $\Delta_{1 \cdot 0.5}$  and  $\Delta_1$ . Because  $\Delta_{1 \cdot 0.5}$  applies a relatively small penalty to the inclusive likelihood ratio when a rate shift occurs at the nested node, its power is consequently more inflated than that of  $\Delta_1$ . Interestingly, the bias of  $\Delta_1$  and  $\Delta_2$  appears to decrease as the magnitude of the rate shift at the nested node increases. The remaining shift statistics,  $\Delta_R$ ,  $\Delta_{NP}$ , and  $\Delta_N$ , appear to overcompensate for rate shifts within the ingroup, such that their power to detect a rate shift at the target node rapidly diminishes with increasing magnitude of rate shifts at the nested node.

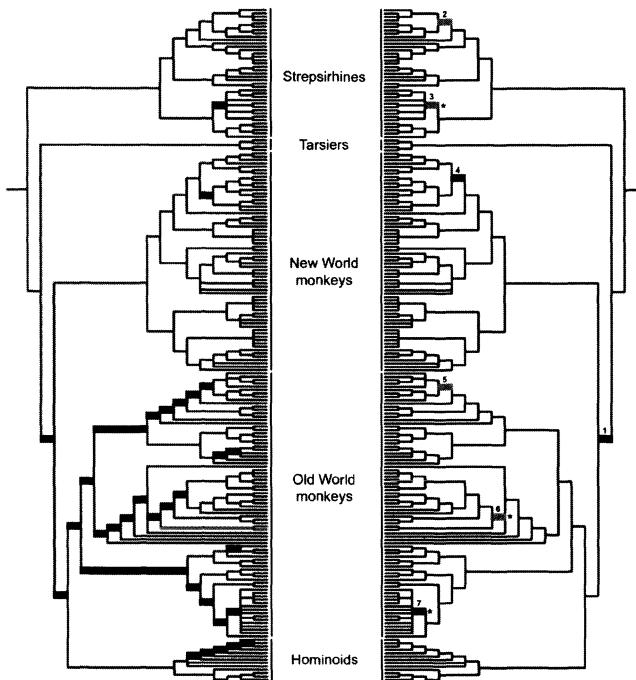
The performance of the various shift statistics under a range of tree sizes is illustrated in Figure 8. These simulations uniformly applied a four-fold diversification rate shift to the internal branch of trees with 100 to 400 tips, a size range reflecting that of supertrees in the literature (e.g., Purvis, 1995; Bininda-Emonds *et al.*, 1999; Wojciechowski *et al.* 2000; Jones *et al.*, 2002; Kennedy and Page, 2002; Salamin *et al.*, 2002; Stoner *et al.*, 2003). Although the power of the shift statistics generally scale with tree size, the increase in power was not realized uniformly by the various tests. The likelihood ratio-based shift statistics,  $\Delta_1$ ,  $\Delta_2$ , and  $\Delta_{1 \cdot 0.5}$ , exhibited the greatest proportional increase in power as tree size increased; by contrast, the power of the  $\Delta_N$  and  $\Delta_{NP}$  statistics was essentially flat across the range of tree sizes simulated, whereas  $\Delta_R$  exhibited an intermediate increase in relative power. Under the range of tree sizes evaluated, the likelihood ratio-based shift statistics consistently exhibited the greatest (and quite similar) absolute

power, accurately identifying a four-fold diversification rate shift ~60% and ~80% of the time in trees with 200 and 400 tips, respectively. Interestingly, although the bias of the  $\Delta_{1*0.5}$  shift statistic increased with tree size, the Type I error rates for the  $\Delta_1$  and  $\Delta_2$  statistics decreased slightly with tree size (not shown). The results discussed above (and illustrated in Figures 7 and 8) were obtained using the three-taxon simulation design; however, we also performed a more elaborate, whole-tree simulation study that allowed rate shifts of various frequency and magnitude to be applied to all the internal branches of simulated trees of various sizes. Results of this more sophisticated study (not shown) were similar qualitatively to those reported for the simpler investigation.

### 4.3 Locating diversification rate shifts in primates

We used the  $\Delta$  shift statistics to locate significant diversification rate shifts in the primate supertree published by Purvis (1995). As in the previous analysis of diversification rate variation using the whole-tree statistics, this analysis is intended for illustrative and comparative purposes only. Accordingly, we made no attempt to account for the effects of phylogenetic error. Results for the two likelihood ratio-based shift statistics,  $\Delta_1$  and  $\Delta_2$ , were obtained using the SYMMETREE program. Polytomies were treated by generating 1000 random resolutions using the size-sensitive ERM taxon-addition algorithm, providing an estimate of the confidence intervals for  $P$ -values associated with each shift statistic. As before, analyses were performed on both the entire primate tree and several of its component clades (e.g., strepsirrhines, New World monkeys, colobines, cercopithecines, and hominoids). After summarizing the findings of our analysis, we describe the methods used in a previous investigation of diversification rate shifts in the primate tree (Purvis *et al.*, 1995), comparing and contrasting the results obtained by these two studies.

Our analysis using the  $\Delta_1$  and  $\Delta_2$  statistics detected seven diversification rate shifts in the primate tree. Significant rate shifts (Figure 9; bold black branches) were detected at the base of haplorhines (along branch 1, the root of a clade comprising New World monkeys, Old World monkeys, and hominoids), within New World monkeys (along branch 4, the root of a clade comprising *Callithrix*, *Cebuella*, *Leontopithecus*, and *Saguinus*), and within Old World monkeys (along branch 7, the root of a clade comprising all *Presbytis* species except *P. entellus*). Additionally, several marginally significant rate shifts (Figure 9; bold gray branches) were detected, including two shifts within the Old World monkey clade (the first along branch 5, the root of a clade comprising *Macaca arctoides*, *M. assamensis*, *M. cyclopis*,



**Figure 9.** Location of inferred diversification rate shifts in the primate supertree of Purvis (1995). The tree at left depicts results from a previous study by Purvis *et al.* (1995) that identified diversification-rate shifts using the relative-cladogenesis statistic (to identify anomalously diverse lineages) coupled with a parsimony optimization scheme: all 32 diversification-rate shifts are shown, including 23 from the simultaneous analysis of the entire tree and an additional nine non-redundant shifts from the analyses of the five component clades. The tree at right depicts diversification rate shifts identified using the  $\Delta$  shift statistics. Results obtained under the two approaches are somewhat correspondent: five of the diversification rate shifts identified by the  $\Delta$  shift statistics are among those identified in the previous study. However, several nested shifts within Old World monkeys (i.e., those involving *Macaca*, *Cercopithecus*, *Presbytis* at branches 5, 6, and 7, respectively) caused a cascade of spurious diversification-rate shifts to be identified at more inclusive nodes throughout the anthropoid clade owing to the trickle-down problem. Bold black branches correspond to significant rate shifts and bold gray branches to marginally significant rate shifts; branches marked with an asterisk identify results involving random resolution of polytomies; numbered branches correspond to clades identified in the text.

*M. fascicularis*, *M. fuscata*, *M. mulatta*, *M. radiata*, *M. sinica*, and *M. thibetana*; and the second shift along branch 6, the root of a clade comprising all *Cercopithecus* species except *C. aethiops* and *C. solatus*) and two within strepsirrhines (the first along branch 2, the root of a clade

comprising *Lemur*, *Hapalemur*, and *Eulemur*; and the second shift along branch 3, the root of a clade comprising *Galago*, *Galagooides*, *Otolemur*, and *Euoticus*). Interestingly, the diversification rate shift along branch 2 was independently identified as a significant radiation in a recent study by Yoder and Yang (in press), which estimated divergence times from several unlinked loci and external fossil calibrations using Bayesian methods.

Several aspects of these findings warrant comment. First, three of the diversification rate shifts were associated with polytomies (i.e., those in *Macaca*, *Presbytis*, and the *Callithrix-Saguinus* clades; Figure 9), demonstrating the applicability of these methods to incompletely resolved (super)trees. Second, in contrast to our previous analysis of diversification rate variation using the whole-tree  $M$  statistics, results obtained using the  $\Delta$  shift statistics were insensitive to the specification of taxonomic scope, returning the same  $P$ -values for the same set of branches regardless of whether the analysis was applied simultaneously to the entire tree or separately to its component clades. Third, the results of the whole-tree  $M$  statistics and  $\Delta$  shift statistics are not perfectly correspondent. Specifically, the  $\Delta$  shift statistics failed to locate significant diversification rate shifts within several clades in which significant among-lineage diversification rate variation had previously been identified by the whole-tree  $M$  statistics. In these cases, diversification rate variation appears to be rather evenly dispersed across the tree such that, although cumulatively significant under the whole-tree  $M$  statistics, it is nevertheless insufficiently concentrated along any one branch (or small number of branches) to constitute a significant diversification rate shift under the  $\Delta$  shift statistics. For example, the topology of the hominoid clade is largely pectinate, indicating significant heterogeneity in diversification rate among its branches. Nevertheless, evaluating the probability of a diversification rate shift along any particular branch is likely to involve a diversity partition of  $1:(N - 1)$  at the inclusive node and  $1:(N - 2)$  at the nested node, which is much more consistent with a trickle down in rates than a local shift in rate under the  $\Delta$  shift statistics. Finally and conversely, significant diversification rate shifts were located within clades for which the whole-tree statistics had previously failed to detect significant among-lineage diversification rate variation. In these cases, diversification rate heterogeneity was largely restricted to a single branch (or small number of branches), constituting a significant local rate shift that was nevertheless below the threshold of detection under the whole-tree  $M$  statistics. For example, the New World monkey clade is, overall, very balanced: diversity partitions at most nodes in this tree involve splits of approximately  $(N / 2):(N / 2)$ . The single prominent exception involves the node at which a rate shift was located (branch 4 in Figure 9), which by itself

was insufficient to cause rejection of the null hypothesis that the whole New World monkey clade diversified under a stochastic ERM branching model.

#### 4.3.1 The relative cladogenesis statistic: potential limitations and comparison to the $\Delta$ shift statistics

The location of diversification rate shifts in the primate supertree was previously studied by Purvis *et al.* (1995) using an approach referred to as the “relative cladogenesis statistic,” originally described in Nee *et al.* (1992, 1994b, 1994b) and subsequently in Harvey and Nee (1993, 1994), Nee and Harvey (1994), and Nee *et al.* (1994a, 1995, 1996). Like the whole-tree statistics described previously, the relative cladogenesis statistic was originally intended to detect significant diversification rate variation among a set of lineages. In contrast to our strictly topology-based whole-tree statistics, however, the relative cladogenesis statistic relies on temporal information to circumscribe the set of lineages involved in the test. That is, given a phylogeny with estimated divergence times, we can arbitrarily draw a line through the tree at some point in the past,  $t_k$ , to identify a set of  $k$  contemporary ancestral lineages. Suppose that these  $k$  ancestral lineages survive to the present and give rise collectively to  $N$  extant descendants, such that the  $i$ th ancestral lineage leaves  $n_i$  extant species, where  $n_i \geq 1$  (because all  $k$  ancestral lineages have survived) and where the  $n_i$  sum to  $N$ . If the  $k$  lineages all diversified at the same rate, then all vectors of descendant species diversities ( $n_1, n_2, n_3, \dots, n_k$ ) are equiprobable<sup>5</sup> (e.g., Nee *et al.*, 1992, 1994b, 1996; Nee and Harvey, 1994; Purvis *et al.*, 1995; Purvis, 1996). This expectation can be used to calculate the probability that one of the ancestral lineages will realize more than  $r$  descendants, given a total of  $N$  species descended from the set of  $k$  ancestral lineages. This probability is given by

$$(10) \quad P = 1 - \frac{\sum_{v=0}^{\infty} (-1)^v \binom{k}{v} \binom{N - rv - 1}{k - 1}}{\binom{N - 1}{k - 1}},$$

<sup>5</sup> Curiously, it is often asserted that the relative cladogenesis test “makes no assumptions about how the clades have been growing” (Nee and Harvey, 1994:1550) and that it “does not depend on any particular model of diversification” (Nee *et al.*, 1996:241; see also Nee and Harvey, 1994; Purvis *et al.*, 1995). Clearly, however, the assumptions entailed by this test—that rates of diversification are equal and independent in all lineages at any given point in time—are those specifying the stochastic ERM random branching model. In fact, equation (12) reduces to equation (1) (which provides the ERM nodal probability) when  $k = 2$  (i.e., for sister-group comparisons where  $N$  descendant species are partitioned among two ancestral sister lineages; e.g., Nee and Harvey, 1994; Nee *et al.*, 1994a, 1995, 1996; Purvis, 1996).

where the summation is for positive  $N - rv - 1$  and where  $N - rv - 1 \geq k - 1$  (Purvis *et al.*, 1995; Nee *et al.*, 1996). A significant result indicates that the clade in question is anomalously diverse and therefore has diversified under a significantly different rate than its contemporaries.

Although originally intended as a test of significant diversification rate variation, the relative cladogenesis statistic was subsequently extended to infer the location of significant diversification rate shifts by Purvis *et al.* (1995). This extension is based on parsimony optimization: if two sister lineages are inferred to be anomalously diverse under the relative cladogenesis statistic, then a significant shift in rate is inferred to have occurred in their common ancestor. Below we consider several potential limitations associated with the attempt to use the relative cladogenesis statistic to locate diversification rate shifts: some limitations are inherent to the method, others pertain more specifically to divergence-time estimates in supertrees. These limitations are illustrated with reference to the analysis of diversification rate shifts in the primate supertree, and compared with the behavior of the  $\Delta$  shift statistics where appropriate.

#### **4.3.1.1 Susceptibility of the relative cladogenesis statistic to arbitrary delineation of test window**

As described above, the relative cladogenesis statistic requires delineation of a “window” within which the test is to be applied. The dimensions of this window include both its temporal depth and its taxonomic breadth. However, circumscription of this window is arbitrary and therefore potentially problematic given that the results inferred from the test are known to be sensitive to the temporal depth (Purvis, 1996) and taxonomic scope specified. Specification of the temporal dimension can be made less arbitrary by sliding the window over the tree from the root to its tips, recalculating the relative cladogenesis statistic at every point in time,  $t_k$ , associated with an increase in  $k$ , the number of the ancestral lineages (where  $k = 2, 3, 4, \dots, (N - 1)$ ). This approach was used by Purvis *et al.* (1995) and has also been implemented in the (now defunct) End-Epi program (Harvey *et al.*, 1996; Rambaut *et al.*, 1997). However, it is considerably more difficult to objectively define (or integrate over) the taxonomic breadth of the comparison, which nevertheless exerts a similarly strong influence on the conclusions obtained. Although the sensitivity of the relative cladogenesis statistic to phylogenetic scope is appropriate when the test is used to detect diversification rate variation, this sample dependency is inappropriate when applied to the problem of locating diversification rate shifts. This problem is manifest in the analysis of diversification rate shifts in the primate supertree: Purvis *et al.* (1995) originally detected 23 significant rate shifts when the

relative cladogenesis statistic was applied to the entire primate tree but subsequently identified an additional nine non-redundant rate shifts when the various component clades were analyzed separately. By contrast, the number of branches identified by the  $\Delta$  shift statistics (and the  $P$ -values of the statistics) were unaffected by the phylogenetic scope of the analysis.

#### 4.3.1.2 Susceptibility of the relative cladogenesis statistic to error in divergence times

Application of the relative cladogenesis test requires reliable estimates of divergence times, which is likely to be problematic for the analysis of supertrees. Error in divergence-time estimates can confound the test by causing misspecification of the appropriate set of ancestral lineages present at the specified  $t_k$ . Although recent methodological and theoretical advances have greatly improved the accuracy of divergence-time estimates derived from the primary analysis of nucleotide sequence data (e.g., Sanderson, 1997, 2002; Rambaut and Bromham, 1998; Thorne *et al.*, 1998; Huelsenbeck *et al.*, 2000a; Yoder and Yang, 2000; Kishino *et al.*, 2001; Thorne and Kishino, 2002), the extent to which these methods can be extended to the estimation of divergence times in supertrees is presently unknown. Close inspection of the primate phylogeny illustrates some of the challenges of estimating divergence times in supertrees, as well as the undesirable consequences of the associated error for inferences of diversification rates that rely on temporal information. We wish to emphasize, however, that our criticisms are not intended to imply that the dates in this particular supertree are exceptionally unreliable; rather, we believe that the level of uncertainty in these data is similar to that in other published supertrees.

Divergence times were estimated for 90 of the 160 nodes in the primate supertree, all of which were derived directly from or calibrated against the primate fossil record. Under the approach used, the divergence time of a clade was equated with the age of the oldest fossil attributed to that lineage. This approach will tend to systematically underestimate the true divergence times of clades in proportion to their degree of incompleteness in the fossil record. For several reasons, the degree to which a lineage is represented in the fossil record is likely to be phylogenetically biased. For example, preservation potential will be influenced by phylogenetically autocorrelated differences in anatomy and demography, and taphonomic factors will be influenced by phylogenetically autocorrelated differences in habitat preference. Consequently, clades will vary in the degree to which their inferred divergence times will be underestimated. The resulting phylogenetically biased error in divergence-time (under)estimation will

induce a corresponding pathological bias for the study of diversification rates: an underestimate in the age of a clade will cause a corresponding overestimate in its inferred rate of diversification.

As expected of a group with a heterogeneous representation in the fossil record, the number of available fossil calibrations varied markedly across the primate supertree: 15 estimates were used to date one node, whereas the divergence times of many others were based on a single estimate. The uncertainty associated with divergence times based on single estimates was approximately and conservatively estimated to have an average error margin  $>\pm 50\%$ , prompting Purvis (1995:413) to reasonably conclude that “not too much reliance should be placed on single estimates.” Nevertheless, several diversification rate shifts in the primate supertree relied on the single date estimates. For example, Purvis *et al.* (1995:331) were appropriately skeptical of the inferred diversification rate shift in the strepsirrhine clade because “the age of the galagid radiation is based on only a single estimate, so it may be inaccurate.” However, this caveat applies equally to several other clades in which diversification rate shifts were detected (e.g., *Cercopithecus*, *Colobus*, *Macaca*, *Presbytis*, *Saguinus*) because they were similarly based on a single (or very few) estimates.

Moreover, several of the nodes based on single (or very few) estimates were used to calibrate other nodes in the primate supertree, causing a cascade of error in both estimation of divergence times and the associated inference of diversification rate shifts. For example, the divergence time of Old World monkey-hominoid clade was estimated by Purvis (1995) at  $27.5 \pm 4.5$  million years ago (Mya) based on two fossils. Independent estimates for the age of this node are typically much older. For example, maximum likelihood estimates based on the entire protein-coding region of the mitochondrial genome calibrated with a more reliable external fossil date (the cetacean-artiodactyl divergence at 53–60 Mya) place this divergence in the range of ~38–68 Mya (Arnason *et al.*, 1998; Yoder and Yang, 2000). The discrepancy in the timing of this divergence is somewhat troubling because it was used to calibrate 32 other nodes within the Old World monkey-hominoid clade (A. Purvis, pers. comm.), in which 28 of the 32 total significant diversification rate shifts were detected by Purvis *et al.* (1995).

Uncertainty in divergence times is not restricted to those nodes based on single estimates: dates in the primate supertree based on multiple estimates also had non-trivial error. For example, Purvis (1995:413) reported significant differences in the proportional error in divergence-time estimates among clades in the primate supertree, which would be expected of a group with phylogenetically biased representation in the fossil record. The highest proportional error was found within cercopithecines, in which fully half of the inferred diversification rate shifts occurred. The extent to which the

acknowledged uncertainty in divergence-time estimates influenced this study of diversification rate shifts in the primate supertree is difficult to ascertain; Purvis *et al.* (1995) acknowledged the presence of error in the divergence times and its potential impact on the analysis but did not attempt to quantify the level of uncertainty or assess the sensitivity of the results to this source of error.

Although the divergence times for the 90 dated nodes are likely to be associated with substantial estimation error, the divergence times for the remaining 70 nodes were not estimated at all, but instead generated deterministically under the assumption of what might be called a “branching clock” (*sensu* Sanmartín *et al.*, 2001). Given a deterministic model of exponential diversification, the divergence time of a given node can be calculated as  $t_d = (t_a)(\ln N_d / \ln N_a)$ , where  $t_a$  and  $t_d$  are the ages of the ancestral and descendant nodes, with  $N_a$  and  $N_d$  species, respectively. Given the countless number of hidden parameters influencing diversification rates, the use of a deterministic branching model (particularly one whose fit to the data is not evaluated) is likely to provide an overly simplistic and potentially problematic solution to the problem of specifying the unknown divergence times. The use of a branching model to specify >40% of the divergence times in the primate supertree is likely to bias inferences of diversification rates (Purvis *et al.*, 1995). Many approaches (including the relative cladogenesis test) invoke stochastic branching models to generate the expected distribution of diversification events against which the observed distribution is compared. However, use of a branching clock essentially involves the model-based generation of the “observations” as well. Although it is difficult to ascertain the accuracy of dates generated with this scheme, there is no reason to expect it to be high: these divergence times combine the considerable uncertainty of those estimated from fossil evidence and/or local clocks (from which they are ultimately calibrated) with a branching clock of uncertain justification.

In summary, scrutiny of the primate supertree highlights the challenges of estimating divergence times in supertrees and reveals how the uncertainty in these data can confound attempts to detect diversification rate shifts using the relative cladogenesis statistic or other temporal tests. Although there is reason for optimism that recent efforts will improve the reliability of divergence-time estimates in supertrees (e.g., Lapointe and Cucumel, 1997; Bryant *et al.*, 2004; Lapointe and Levasseur, 2004; Vos and Mooers, 2004), the ability of these methods to provide sufficiently accurate temporal information has yet to be demonstrated. By contrast, because they effectively ignore temporal information, the topology-based  $\Delta$  shift statistics provide a more reliable means with which to infer diversification rate shifts in supertrees.

#### 4.3.1.3 Susceptibility of the relative cladogenesis statistic to the trickle-down problem

In addition to pioneering the development of methods for locating significant diversification rate shifts, Purvis *et al.* (1995:331) were also among the first authors to recognize the potentially confounding influence of what we have termed the trickle-down problem, raising the caveat that any “result must be interpreted cautiously because radiations are not independent: if a given clade is a significant radiation, more inclusive clades will tend to be.” In other words, significant diversification rate shifts at more nested nodes will lead to the identification of spurious diversification rate shifts at more inclusive nodes under their proposed parsimony optimization scheme. Results from the primate analysis provide compelling empirical evidence of the susceptibility of the relative cladogenesis statistic to the trickle-down problem. Despite the aforementioned criticisms, the relative cladogenesis statistic identified several diversification rate shifts also indicated by the likelihood ratio-based  $\Delta$  shift statistics (e.g., shifts located within *Galago*, *Macaca*, *Cercopithecus*, and *Presbytis* at branches 3, 5, 6, and 7, respectively; Figure 9). However, diversification rate shifts detected at relatively nested nodes within the Old World monkey clade (those within *Macaca*, *Cercopithecus*, and *Presbytis* at branches 5, 6, and 7, respectively; Figure 9) caused a trickle-down of diversification rate shifts to be inferred at more inclusive nodes under the relative cladogenesis test. Accordingly, the demonstrable susceptibility of the relative cladogenesis statistic to the trickle-down problem suggests that this test is more appropriately restricted to the inference of diversification rate variation.

## 5. Discussion

### 5.1 Implementation and accommodation of phylogenetic uncertainty

The methods described in this chapter have been implemented in the computer program, SYMMETREE. Executables have been compiled for Macintosh (OS 9 and OS X), Windows, and UNIX operating systems, which are freely available at <http://www.phylodiversity.net/brian/> or <http://www.kchan.org>, or by emailing either of these authors directly.

Methods for detecting diversification rate variation have typically required strictly dichotomous phylogenies; given the empirical reality of polytomies, this limitation has proven to be a serious impediment to their

application. Accordingly, an important feature of SYMMETREE is its facility to deal with incompletely resolved trees. The program recognizes two types of soft polytomies that require different analytical approaches: “collapsed” polytomies, which are caused by internal branches of zero length; and, “consensus” polytomies, which stem from conflict among a set of equally optimal (super)tree estimates. Collapsed polytomies are addressed by randomly (and repeatedly) generating dichotomous solutions using one of several alternative random taxon-addition algorithms. However, this procedure is inappropriate for consensus polytomies. Although such a polytomy might stem from conflict among a small set of source trees, it might nevertheless be consistent with a much larger set of (randomly resolved) binary trees. Accordingly, only those resolutions of a consensus polytomy that belong to the set of conflicting trees should be considered, which can be accomplished by means of a batch-processing option that sequentially analyzes each tree belonging to the set of conflicting trees. For both collapsed and consensus polytomies, the appropriate test can be applied to each tree within the set of (randomly resolved or equally optimal) trees to provide an estimate of the confidence intervals on the inference being made.

More generally, polytomies can be viewed as a manifestation of phylogenetic uncertainty. Although often acknowledged as a crucial assumption, the effect of phylogenetic error on inferences of diversification rate is seldom explicitly taken into account (but see, for example, Sanderson and Wojciechowski, 1996; Baldwin and Sanderson, 1998). In theory, it would be straightforward to assess the confidence interval on an inference by batch processing the bootstrap profile (and/or the posterior probability distribution) of study trees. Although this approach is viable for trees derived from primary analyses (i.e., conventional analysis of the primary character data), supertree estimation methods present a special challenge in this respect because there is currently no comparable means of estimating topological uncertainty in supertrees. Clearly, this area requires further development (Ronquist *et al.*, 2004; Moore *et al.*, in prep.).

## 5.2 Extensions, limitations, and applications

The methods described in this chapter are intended to answer two general questions. Have the branches of this tree experienced differential diversification rates? And, if so, on which branches have those shifts in rate occurred? Accordingly, these methods should find useful application to a range of problems (outlined below) but, of course, will be ill-suited to the investigation of other equally valid and interesting evolutionary questions. For example, we might want to estimate parameters associated with the diversification process (e.g., speciation and extinction rates) or test whether

diversification rates have changed significantly through time. These questions require information on the (relative or absolute) timing of diversification events, and so will necessarily involve the use of temporal methods (e.g., Harvey *et al.*, 1991, 1994a, b; Nee *et al.*, 1992, 1994a, b; Harvey and Nee, 1993, 1994; Kubo and Iwasa, 1995; Paradis, 1997, 1998b; Nee, 2001). In such applications of temporal methods, however, it is first necessary to establish that there has not been significant among-lineage diversification rate variation within the study phylogeny. This requirement can readily be established (or disconfirmed) with our whole-tree tests for diversification rate variation, again emphasizing the inherent complementarity of temporal- and topology-based methods.

Other questions might be profitably addressed by extending the whole-tree methods described herein. For example, we might want to know if shifts in diversification rate are correlated with changes in some other variable (e.g., the origin of morphological or behavioral novelties, ecological associations, or biogeographic events). Topology-based approaches to this problem are available but typically involve replicated sister-group comparisons (e.g., Slowinski and Guyer, 1993; Nee *et al.*, 1996; Barraclough *et al.*, 1998; Goudet, 1999; Simms and McConway, 2003) that incorporate relatively limited phylogenetic information (e.g., Sanderson and Donoghue, 1994, 1996). As has been demonstrated for other diversification rate problems, the power to detect correlates of shifts in diversification rate is likely to be substantially enhanced by incorporating information from more of the tree. We are currently working to extend the methods described in this chapter to provide a whole-tree approach to this problem.

In addition to addressing other types of questions, the whole-tree methods described in this chapter might also be profitably extended to incorporate additional sources of information. Although the whole-tree methods currently utilize exclusively topological information on the distribution of species diversity, they could readily be generalized to incorporate temporal information on the distribution of waiting times between diversification events. It is conceivable that the inclusion of divergence-time estimates, when available and appropriate to the hypothesis of interest, could further enhance the power to detect the presence and locate the position of significant shifts in diversification rate.

Future elaborations notwithstanding, the whole-tree methods presented here have immediate implications for a range of data-exploration and hypothesis-testing scenarios associated with the study of diversification rates. Whole-tree surveys for significant diversification rate variation could provide an effective discovery method for generating causal hypotheses of factors that have caused, are caused by, or are correlated with differential diversification rates. For example, the discovery that diversification rate

variation is often associated with plant clades that are polymorphic for growth form (i.e., woody / herbaceous) might lead us to hypothesize that shifts in growth form are affecting diversification rates (e.g., Eriksson and Bremer, 1991, 1992; Bremer and Eriksson, 1992; Judd *et al.*, 1994; Ricklefs and Renner, 1994; Tiffney and Mazer, 1995; Dodd *et al.*, 1999). The ability to detect clades with significant diversification rate variation and/or diversification rate shifts will also help identify the data relevant to studies of phenomena that are hypothesized to be correlated with differential diversification rates. For example, application of the whole-tree tests could identify the data necessary to evaluate the hypothesized correlation between rates of nucleotide substitution and rates of cladogenesis (e.g., Mindell *et al.*, 1989; Barraclough *et al.*, 1996; Savolainen and Goudet, 1998; Barraclough and Savolainen, 2001; Jobson and Albert, 2002). Additionally, the whole-tree tests could provide more powerful tools for studies that seek to assess the empirical prevalence of diversification rate variation (e.g., Guyer and Slowinski, 1991, 1993; Heard, 1992; Mooers, 1995). Finally, several evolutionary processes could entail hypotheses that predict multiple diversification rate shifts dispersed throughout whole clades, rather than single shifts concentrated at particular nodes. These processes include the effect of various co-evolutionary associations on rates of diversification (e.g., the reciprocal radiations predicted for some insect / plant associations; Farrell, 1998; Farrell and Mitter, 1998; Kelly and Farrell, 1998) and the effect of relative refractory periods associated with “age-biased cladogenesis” (Hey 1992; Harvey and Nee, 1993; Losos and Adler, 1995; Chan and Moore, 1999).

The foregoing discussion suggests that several different evolutionary questions — associated with detecting significant diversification rate variation or locating diversification rate shifts — might be effectively addressed by the separate application of either the whole-tree  $M$  statistics or the  $\Delta$  shift statistics, respectively. However, both sets of methods could be applied in concert to address additional evolutionary questions. For instance, the combined application of the  $M$  and  $\Delta$  statistics might be used to explore the empirical prevalence of different models of cladogenesis (Figure 10). As demonstrated in the primate analyses, results obtained under the whole-tree  $M$  statistics and the  $\Delta$  shift statistics will not always be perfectly correspondent. That is, the whole-tree  $M$  statistics might occasionally detect significant diversification rate variation within clades for which the  $\Delta$  shift statistics subsequently fail to locate any significant diversification rate shifts. Conversely, significant diversification rate shifts might sometimes be identified within clades for which the  $M$  statistics fail to detect significant among-lineage diversification rate variation. The former scenario will arise when diversification rate variation is rather evenly dispersed across the tree

|                                                     |             | Diversification Rate Shift<br>( $\Delta$ statistics) |         |
|-----------------------------------------------------|-------------|------------------------------------------------------|---------|
|                                                     |             | significant                                          | n.s.    |
| Diversification Rate Variation<br>( $M$ statistics) | significant | mixed                                                | gradual |
|                                                     | n.s.        | punctuated                                           | ERM     |

*Figure 10.* Exploiting discord in results obtained under the whole-tree  $M$  statistics and the  $\Delta$  shift statistics to explore modes of diversification. The lower right cell indicates stochastically homogeneous (ERM) diversification rates, whereas the other three scenarios involve diversification-rate heterogeneity consistent with either gradual, punctuated, or mixed evolutionary models of cladogenesis.

(see Section 4.3); such a relatively uniform phylogenetic distribution of diversification rate change is consistent with a gradual evolutionary model of cladogenesis. By contrast, the latter scenario entails a local concentration of diversification rate heterogeneity along a single branch (or small number of branches) that is below the threshold of detection under the whole-tree  $M$  statistics. This relatively sporadic phylogenetic distribution of diversification rate change is consistent with a punctuated evolutionary model of cladogenesis. Thus, discord in the results obtained under the whole-tree  $M$  statistics or the  $\Delta$  shift statistics can be usefully exploited to tease apart modes of diversification rate heterogeneity.

In conclusion, we are optimistic that the methods described in this chapter should enable a range of evolutionary questions to be addressed when reliable temporal information is either unavailable or inappropriate to the problem at hand.

## Acknowledgements

We are grateful to Olaf Bininda-Emonds for inviting us to contribute to this volume on supertrees, and for his Buddha-like patience while enduring its elephantine gestation. We are indebted particularly to Mary Moore for assistance with the illustrations, and to Michael Sanderson and Mary Moore for offering perceptive comments on earlier drafts of this chapter. Thanks are also due to Junhyong Kim for insightful discussions on rate estimation, and to Anne Yoder for helpful discussions of divergence-time estimation in

primates and for providing access to preprints. Exceptionally helpful reviews were provided by Olaf Bininda-Emonds, Peter Mayhew, and Andy Purvis. BRM and KC wish to thank Simon Levin for hosting BRM in the Levin Lab during the course of this project and for his perennial guidance and support of their collaborative endeavors. Although this paper has benefited greatly from the suggestions of the aforementioned readers, we accept sole responsibility for any remaining errors that it might contain. Funding for this research was made possible through Natural Science and Engineering Research Council of Canada (NSERC) postgraduate scholarships to BRM and KC and through a Deep Time RCN graduate training award to BRM. This work was completed while KC was at the Department of Ecology and Evolutionary Biology, Princeton University.

## References

- AGAPOW, P.-M. AND PURVIS, A. 2002. Power of eight tree shape statistics to detect nonrandom diversification: A comparison of two models of cladogenesis. *Systematic Biology* 51:866–872.
- ARNASON, U., GULLBERG, A., AND JANKE, A. 1998. Molecular timing of primate divergences as estimated by two non-primate calibration points. *Journal of Molecular Evolution* 47:718–727.
- BALDWIN B. G. AND SANDERSON, M. J. 1998. Age and rate of diversification of the Hawaiian silversword alliance (Compositae). *Proceedings of the National Academy of Sciences of the United States of America* 95:9402–9406.
- BARRACLOUGH, T. G., HARVEY, P. H., AND NEE, S. 1996. Rate of *rbcL* gene sequence evolution and species diversification in flowering plants (angiosperms). *Proceedings of the Royal Society of London B* 263:589–591.
- BARRACLOUGH, T. G., NEE, S., AND HARVEY, P. H. 1998. Sister-group analysis in identifying correlates of diversification: comment. *Evolutionary Ecology* 12:751–754.
- BARRACLOUGH, T. G. AND NEE, S. 2001. Phylogenetics and speciation. *Trends in Ecology and Evolution* 16:391–399.
- BARRACLOUGH, T. G. AND SAVOLAINEN, V. 2001. Evolutionary rates and species diversity in flowering plants. *Evolution* 55:677–683.
- BININDA-EMONDS, O. R. P., GITTLEMAN, J. L., AND PURVIS, A. 1999. Building large trees by combining phylogenetic information: a complete phylogeny of the extant Carnivora (Mammalia). *Biological Reviews* 74:143–175.
- BREMER, B. AND ERIKSSON, O. 1992. Evolution of fruit characters and dispersal modes in the tropical family Rubiaceae. *Biological Journal of the Linnean Society* 47:79–95.
- BRYANT, D., SEMPLE, C., AND STEEL, M. 2004. Supertree methods for ancestral divergence dates and other applications. In O. R. P. Bininda-Emonds (ed). *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 129–150. Kluwer Academic, Dordrecht, The Netherlands.
- CHAN, K. M. A. AND MOORE, B. R. 1999. Accounting for mode of speciation increases power and realism of tests of phylogenetic asymmetry. *American Naturalist* 153:332–346.
- CHAN, K. M. A. AND MOORE, B. R. 2002. Whole-tree methods for detecting differential diversification rates. *Systematic Biology* 51:855–865.

- COLLESS, D. H. 1982. Review of Phylogenetics: The Theory and Practice of Phylogenetic Systematics, by E. O. Wiley. *Systematic Zoology* 31:100–104.
- DODD, M. E., SILVERTOWN, J., AND CHASE, M. W. 1999. Phylogenetic analysis of trait evolution and species diversity variation among angiosperm families. *Evolution* 53:732–744.
- DONOGHUE, M. J. AND ACKERLY, D. D. 1996. Phylogenetic uncertainties and sensitivity analyses in comparative biology. *Philosophical Transactions of the Royal Society of London B* 351:1241–1249.
- DOYLE, J. A. AND DONOGHUE, M. J. 1993. Phylogenies and angiosperm diversification. *Paleobiology* 19:141–167.
- EDGINGTON, E. S. 1972a. An additive method for combining probability values from independent experiments. *Journal of Psychology* 80:351–363.
- EDGINGTON, E. S. 1972b. A normal curve method for combining probability values from independent experiments. *Journal of Psychology* 82:85–89.
- EDGINGTON, E. S. AND HALLER, O. 1984. Combining probabilities from discrete probability distributions. *Educational and Psychological Measurement* 44: 265–274.
- ERIKSSON, O. AND BREMER, B. 1991. Fruit characteristics, life forms, and species richness in the plant family Rubiaceae. *American Naturalist* 138:751–761.
- ERIKSSON, O. AND BREMER, B. 1992. Pollination systems, dispersal modes, life forms, and diversification rates in angiosperm families. *Evolution* 46:258–256.
- FARRELL, B. D. 1998. “Inordinate fondness” explained: why are there so many beetles? *Science* 281:555–559.
- FARRELL, B. D. AND MITTER, C. 1998. The timing of insect/plant diversification: might *Tetraopes* (Coleoptera: Cerambycidae) and *Asclepias* (Asclepiadaceae) have co-evolved? *Biological Journal of the Linnean Society* 63:553–577.
- FELSENSTEIN, J. 1988. Phylogenies from molecular sequences: Inference and reliability. *Annual Review of Genetics* 22:521–565.
- FELSENSTEIN, J. 1989. PHYLIP – Phylogeny Inference Package (Version 3.2). *Cladistics* 5:164–166. (<http://evolution.genetics.washington.edu/phylip.html>)
- FISHER, R. A. 1932. *Statistical Methods for Research Workers*. 4th edition. Oliver and Boyd, Edinburgh.
- FURNAS, G. W. 1984. The generation of random, binary unordered trees. *Journal of Classification* 1:187–233.
- FUSCO, G. AND CRONK, Q. C. B. 1995. A new method for evaluating the shape of large phylogenies. *Journal of Theoretical Biology* 175:235–243.
- GOUDET, J. 1999. An improved procedure for testing the effects of key innovations on rate of speciation. *American Naturalist* 153:549–555.
- GOULD, S. J., RAUP, D. M., SEPOWSKI, J. J., SCHOPF, T. J. M., AND SIMBERLOFF, D. S. 1977. The shape of evolution: a comparison of real and random clades. *Paleobiology* 3:23–40.
- GUYER, C. AND SLOWINSKI, J. B. 1991. Comparisons of observed phylogenetic topologies with null expectations among three monophyletic lineages. *Evolution* 45:340–350.
- GUYER, C. AND SLOWINSKI, J. B. 1993. Adaptive radiations and the topology of large phylogenies. *Evolution* 47:253–263.
- HARRIS, T. E. 1964. *The Theory of Branching Processes*. Springer-Verlag, Berlin.
- HARVEY, P. H., NEE, S., MOOERS, A. Ø., AND PARTRIDGE, L. 1991. These hierarchical views of life: phylogenies and metapopulations. In R. J. Berry, T. J. Cranford, and G. M. Hewitt (eds), *Genes in Ecology*, pp. 123–137. Blackwell Scientific, Oxford.
- HARVEY, P. H. AND NEE, S. 1993. New uses for new phylogenies. *European Review* 1:11–19.

- HARVEY, P. H. AND NEE, S. 1994. Comparing real with expected patterns from molecular phylogenies. In P. Eggleton and R. I. Vane-Wright (eds), *Phylogenetics and Ecology*, pp. 219–231. Academic Press, London.
- HARVEY, P. H., HOLMES, E. C., MOOERS, A. Ø., AND NEE, S. 1994a. Inferring evolutionary processes from molecular phylogenies. In R. W. Scotland, D. J. Siebert, and D. M. Williams (eds), *Models in Phylogeny Reconstruction*, pp. 313–333. Clarendon Press, Oxford.
- HARVEY, P. H., MAY, R. M., AND NEE, S. 1994b. Phylogenies without fossils. *Evolution* 48:523–529.
- HARVEY, P. H., RAMBAUT, A., AND NEE, S. 1996. New computer packages for analysing phylogenetic tree structure. In J. Colbert and R. Barbault (eds), *Aspects of the Genesis and Maintenance of Biological Diversity*, pp. 60–68. Oxford University Press, Oxford.
- HARDING, E. F. 1971. The probabilities of rooted tree-shapes generated by random bifurcation. *Advances in Applied Probability* 3:44–77.
- HEARD, S. B. 1992. Patterns in tree balance among cladistic, phenetic, and randomly generated phylogenetic trees. *Evolution* 46:1818–1826.
- HENNIG, W. 1966. *Phylogenetic Systematics*. University of Illinois Press, Urbana, Illinois.
- HEY, J. 1992. Using phylogenetic trees to study speciation and extinction. *Evolution* 46:627–640.
- HUELSENBECK, J. P., LARGET, B., AND SWOFFORD, D. 2000a. A compound Poisson process for relaxing the molecular clock. *Genetics* 154:1879–1892.
- HUELSENBECK, J. P., RANNALA, B., AND MASLY, J. P. 2000b. Accommodating phylogenetic uncertainty in evolutionary studies. *Science* 288:2349–2350.
- HULBERT, R. C. 1993. Taxonomic evolution in North American Neogene horses (subfamily Equinae): the rise and fall of an adaptive radiation. *Paleobiology* 19:216–234.
- JENSEN, J. S. 1990. Plausibility and testability: Assessing the consequences of evolutionary innovation. In M. H. Nitecki (ed.), *Evolutionary Innovations*, pp. 171–190. University of Chicago Press, Chicago.
- JOBSON, R. W. AND ALBERT, V. A. 2002. Molecular rates parallel diversification contrasts between carnivorous plant sister lineages. *Cladistics* 18:127–136.
- JONES, K. E., PURVIS, A., MACLARNON, A., BININDA-EMMONDS, O. R. P., AND SIMMONS, N. B. 2002. A phylogenetic supertree of the bats (Mammalia: Chiroptera). *Biological Reviews* 77:223–259.
- JUDD, W. S., SANDERS, R. W., AND DONOGHUE, M. J. 1994. Angiosperm family pairs: preliminary phylogenetic analyses. *Harvard Papers in Botany* 1:1–51.
- KELLEY S. T. AND FARRELL, B. D. 1998. Is specialization a dead end? The phylogeny of host use in *Dendroctonus* bark beetles (Scolytidae). *Evolution* 52:1731–1743.
- KENDALL, D. G. 1948. On the generalized birth-and-death process. *Annals of Mathematical Statistics* 19:1–15.
- KENNEDY, M. AND PAGE, R. D. M. 2002. Seabird supertrees: combining partial estimates of procellariiform phylogeny. *The Auk* 119:88–108.
- KIRKPATRICK, M. AND SLATKIN, M. 1993. Searching for evolutionary patterns in the shape of a phylogenetic tree. *Evolution* 47:1171–1181.
- KISHINO, H., THORNE, J. L., AND BRUNO, W. J. 2001. Performance of divergence time estimation methods under a probabilistic model of rate evolution. *Molecular Biology and Evolution* 18: 352–361.
- KUBO, T. AND IWASA, Y. 1995. Inferring rates of branching and extinction from molecular phylogenies. *Evolution* 49:694–704.

- LAPOINTE, F.-J. AND CUCUMEL, G. 1997. The average consensus procedure: combination of weighted trees containing identical or overlapping sets of taxa. *Systematic Biology* 46:306–312.
- LAPOINTE, F.-J. AND LEVASSEUR, C. 2004. Everything you always wanted to know about the average consensus, and more. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 87–105. Kluwer Academic, Dordrecht, the Netherlands.
- LEE, M. S. Y. 1999. Molecular clock calibrations and Metazoan divergence dates. *Journal of Molecular Evolution* 49:385–391.
- LOSOS, J. B. AND ADLER, F. R. 1995. Stumped by trees? A generalized null model for patterns of organismal diversity. *American Naturalist* 145:329–342.
- MADDISON, W. P. 1989. Reconstructing character evolution on polytomous cladograms. *Cladistics* 5:365–377.
- MAGALLÓN, S. AND SANDERSON, M. J. 2001. Absolute diversification rates in angiosperm clades. *Evolution* 55:1762–1780.
- MCKENZIE, A. AND STEEL, M. 2000. Distributions of cherries for two models of trees. *Mathematical Biosciences* 164:81–92.
- MINDELL, D. P., SITES, J. W., JR., AND GRAUR, D. 1989. Speciational evolution: a phylogenetic test with allozymes in *Sceloporus* (Reptilia). *Cladistics* 5:49–61.
- MOOERS, A. Ø. 1995. Tree balance and tree completeness. *Evolution* 49:379–384.
- MOOERS, A. Ø. AND HEARD, S. B. 1997. Inferring evolutionary process from phylogenetic tree shape. *Quarterly Review of Biology* 72:31–54.
- NEE, S. 2001. Inferring speciation rates from phylogenies. *Evolution* 55:661–668.
- NEE, S., MOOERS, A. Ø., AND HARVEY, P. H. 1992. Tempo and mode of evolution revealed from molecular phylogenies. *Proceedings of the National Academy of Sciences of the United States of America* 89:8322–8326.
- NEE, S., R. AND HARVEY, P. H. 1994. Getting to the root of flowering plant diversity. *Science* 264:1549–1550.
- NEE, S., HOLMES, E. C., MAY, R. M., AND HARVEY, P. H. 1994a. Extinction rates can be estimated from molecular phylogenies. *Philosophical Transactions of the Royal Society of London B* 344:77–82.
- NEE, S., MAY, R. M., AND HARVEY, P. H. 1994b. The reconstructed evolutionary process. *Philosophical Transactions of the Royal Society of London B* 344:305–311.
- NEE, S., HOLMES, E. C., MAY, R. M., AND HARVEY, P. H. 1995. Estimating extinction from molecular phylogenies. In J. H. Lawton and R. M. May (eds), *Extinction Rates*, pp. 164–182. Oxford University Press, Oxford.
- NEE, S., BARRACLOUGH, T. G., AND HARVEY, P. H. 1996. Temporal changes in biodiversity: detecting patterns and identifying causes. In K. J. Gaston (ed.), *Biodiversity: a Biology of Numbers and Differences*, pp. 230–252. Blackwell Science, Oxford.
- PAGE, R. D. M. 1993. On describing the shape of rooted and unrooted trees. *Cladistics* 9:93–99.
- PARADIS, E. 1997. Assessing temporal variations in diversification rates from phylogenies: Estimation and hypothesis testing. *Proceedings of the Royal Society of London B* 264:1141–1147.
- PARADIS, E. 1998a. Detecting shifts in diversification rates without fossils. *American Naturalist* 152:176–187.
- PARADIS, E. 1998b. Testing for constant diversification rates using molecular phylogenies: a general approach based on statistical tests for goodness of fit. *Molecular Biology and Evolution* 15:476–479.

- PURVIS, A. 1995. A composite estimate of primate phylogeny. *Philosophical Transactions of the Royal Society of London B* 348:405–421.
- PURVIS, A. 1996. Using interspecies phylogenies to test macroevolutionary hypotheses. In K. J. Gaston (ed.), *Biodiversity: a Biology of Numbers and Differences*, pp. 151–168. Blackwell Science, Oxford.
- PURVIS, A., NEE, S., AND HARVEY, P. H. 1995. Macroevolutionary inferences from primate phylogeny. *Proceedings of the Royal Society of London B* 260:329–333.
- PURVIS, A., KATZOURAKIS, A., AND AGAPOW, P.-M. 2002. Evaluating phylogenetic tree shape: two modifications to Fusco and Cronk's method. *Journal of Theoretical Biology* 214:99–103.
- PYBUS, O. G. AND HARVEY, P. H. 2000. Testing macro-evolutionary models using incomplete molecular phylogenies. *Proceedings of the Royal Society of London B* 267:2267–2272.
- PYBUS, O. G., RAMBAUT, A., HOLMES, E. C., AND HARVEY, P. H. 2002. New inferences from tree shape: numbers of missing taxa and population growth rates. *Systematic Biology* 51:881–888.
- RAMBAUT, A., HARVEY, P. H., AND NEE, S. 1997. End-Epi: an application for inferring phylogenetic and population dynamical processes from molecular sequences. *Computer Applications in the Biosciences* 13:303–306.
- RAMBAUT, A. AND BROMHAM, L. 1998. Estimating divergence dates from molecular sequences. *Molecular Biology and Evolution* 15:442–448.
- RAUP, D. M., GOULD, S. J., SCHOPF, T. J. M., AND SIMBERLOFF, D. S. 1973. Stochastic models of phylogeny and the evolution of diversity. *Journal of Geology* 81:525–542.
- RICKLEFS, R. E. AND RENNER, S. S. 1994. Species richness within families of flowering plants. *Evolution* 48:1619–1636.
- ROGERS, J. S. 1993. Response of Colless's tree imbalance to number of terminal taxa. *Systematic Biology* 42:102–105.
- ROGERS, J. S. 1994. Central moments and probability distribution of Colless' coefficient of tree imbalance. *Evolution* 48:2026–2036.
- ROGERS, J. S. 1996. Central moments and probability distributions of three measures of phylogenetic tree imbalance. *Systematic Biology* 45:99–110.
- RONQUIST, F., HUELSENBECK, J. P., AND BRITTON, T. 2004. Bayesian supertrees. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 193–224. Kluwer Academic, Dordrecht, the Netherlands.
- SALAMIN, N., HODKINSON, T. R., AND SAVOLAINEN, V. 2002. Building supertrees: an empirical assessment using the grass family (Poaceae). *Systematic Biology* 51:136–150.
- SANDERSON, M. J. 1994. Reconstructing the history of evolutionary processes using maximum likelihood. In D. M. Fambrough (ed.), *Molecular Evolution of Physiological Processes*, Society of General Physiologists Series 49:13–26. Rockefeller University Press, New York.
- SANDERSON, M. J. 1997. A non-parametric approach to estimating divergence times in the absence of rate constancy. *Molecular Biology and Evolution* 14:1218–1231.
- SANDERSON, M. J. 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Molecular Biology and Evolution* 19:101–109.
- SANDERSON, M. J. AND BHARATHAN, G. 1993. Does cladistic information affect inferences about branching rates? *Systematic Biology* 42:1–17.
- SANDERSON, M. J. AND DONOGHUE, M. J. 1994. Shifts in diversification rate with the origin of angiosperms. *Science* 264:1590–1593.
- SANDERSON, M. J. AND DONOGHUE, M. J. 1996. Reconstructing shifts in diversification rates on phylogenetic trees. *Trends in Ecology and Evolution* 11:15–20.

- SANDERSON, M. J. AND WOJCIECHOWSKI, M. F. 1996. Diversification rates in a temperate legume clade: are there “so many species” of *Astragalus* (Fabaceae)? *American Journal of Botany* 83:1488–1502.
- SANMARTÍN, I., ENGHOF, H., AND RONQUIST, F. 2001. Patterns of animal dispersal, vicariance and diversification in the Holarctic. *Biological Journal of the Linnean Society* 73:345–390.
- SARICH, V. AND WILSON, A. C. 1967. Rates of albumin evolution in primates. *Proceedings of the National Academy of Sciences of the United States of America* 58:142–148.
- SAVOLAINEN, V. AND GOUDET, J. 1998. Rate of gene sequence evolution and species diversification in flowering plants: a re-evaluation. *Proceedings of the Royal Society of London B* 265:603–607.
- SHAO, K.-T. AND SOKAL, R. R. 1990. Tree balance. *Systematic Zoology* 39:266–276.
- SIMMS, H. J. AND MCCONWAY, K. J. 2003. Nonstochastic variation of species-level diversification rates within angiosperms. *Evolution* 57:460–479.
- SLOWINSKI, J. B. 1990. Probabilities of  $n$ -trees under two models: Demonstration that asymmetrical interior nodes are not improbable. *Systematic Zoology* 39:89–94.
- SLOWINSKI, J. B. AND GUYER, C. 1989a. Testing the stochasticity of patterns of organismal diversity: an improved null model. *American Naturalist* 134:907–921.
- SLOWINSKI, J. B. AND GUYER, C. 1989b. Testing null models in questions of evolutionary success. *Systematic Zoology* 38:189–191.
- SLOWINSKI, J. B. AND GUYER, C. 1993. Testing whether certain traits have caused amplified diversification: an improved method based on a model of random speciation and extinction. *American Naturalist* 142:1019–1024.
- STANLEY, S. M. 1979. *Macroevolution: Pattern and Process*. W. H. Freeman, San Francisco.
- STONE, J. AND REPKA, J. 1998. Using a nonrecursive formula to determine cladogram probabilities. *Systematic Biology* 47:617–624.
- STONER, C. J., BININDA-EMONDS, O. R. P., AND CARO, T. 2003. The adaptive significance of coloration in lagomorphs. *Biological Journal of the Linnean Society* 79:309–328.
- TAKEZAKI, N., RZHETSKY, A., AND NEI, M. 1995. Phylogenetic test of the molecular clock and linearized trees. *Molecular Biology and Evolution* 12:823–833.
- THORNE, J. L., KISHINO, H., AND PAINTER, I. S. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution* 15:1647–1657.
- THORNE, J. L. AND KISHINO, H. 2002. Divergence time and evolutionary rate estimation with multilocus data. *Systematic Biology* 51:689–702.
- TIFFNEY, B. H. AND MAZER, S. J. 1995. Angiosperm growth habit, dispersal and diversification reconsidered. *Evolutionary Ecology* 9:93–117.
- VOS, R. A. AND MOOERS, A. Ø. 2004. Reconstructing divergence times for supertrees. In O. R. P. Bininda-Emonds (ed.), *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 281–299. Kluwer Academic, Dordrecht, the Netherlands.
- WALLIS, W. A. 1942. Compounding probabilities from independent significance tests. *Econometrica* 10:229–248.
- WILEY, E. O. 1981. *Phylogenetics: the Theory and Practice of Phylogenetic Systematics*. Wiley and Sons, New York.
- WOJCIECHOWSKI, M. F., SANDERSON, M. J., STEEL, K. P., AND LISTON, A. 2000. Molecular phylogeny of the “temperate herbaceous tribes” of papilionoid legumes: a supertree approach. In P. Herendeen and A. Bruneau (eds), *Advances in Legume Systematics* 9:277–298. Royal Botanic Garden, Kew.
- WU, C.-I. AND LI, W.-H. 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. *Proceedings of the National Academy of Sciences of the United States of America* 82:1741–1745.

- YODER, A. D. AND YANG, Z. H. 2000. Estimation of speciation dates using local molecular clocks. *Molecular Biology and Evolution* 17:1081–1090.
- YODER, A. D. AND YANG, Z. H. In press. Divergence dates for Malagasy lemurs estimated from multiple gene loci: fit with climatological events and speciation models. *Molecular Ecology*
- YULE, G. U. 1924. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis. *Philosophical Transactions of the Royal Society of London B* 213:21–87.

## Taxon index

At least to a limited extent, entries in this index take the form of an indented taxonomic hierarchy, where some taxon names are listed as subheadings under a more inclusive name.

- Angiospermae, 11, 77, 79–80, 461–467,  
    472–481
- Apiales, 29
- Arabidopsis thaliana*, 78
- Asteraceae, 314, 476
- Cyperaceae, 465, 476
- Disperis, 465
- Ecdeiocoleaceae, 476
- Fabaceae, 29
- Fagaceae, 465
  - Lithocarpus*, 29
- Joinvilleaceae, 476
- Juncaceae, 465
- Lycopersicon*, 20
- Orchidaceae, 478
- Poaceae, 11, 29, 360, 370, 461–463,  
    465–478, 480–481
- Arachnida, 251–254
  - Araneoclada, 251–253
  - Araneomorphae, 251–252
  - Austrochilidae, 251
  - Neocribellatae, 251–253
- Archaea, 314
- Aves, 250, 257–259, 449
  - Anas*, 259
  - Gallus*, 250, 258–259
- Procellariiformes, 29, 181, 185–186,  
    370
- Struthio*, 259
- bacteria, 29, 304, 314, 370, 450
  - Acholeplasmataceae, 314
  - Bifidobacteriales, 314
  - Cyanobacteria, 314
  - Flexibacteraceae, 314
  - Spirochaetes, 314
- Dinosauria, 29, 276, 372, 413, 455
- Drosophila*, 113–117, 206, 209
- Mammalia, 9, 29, 151, 250, 254, 256,  
    259, 267–269, 273, 277, 363,  
    370–371, 373, 379, 382, 413–415,  
    430, 440, 442, 447, 451, 455, 467
- Afrotheria, 373
- Artiodactyla, 11, 29, 269, 411–431,  
    520
  - Bovidae, 420–426, 429, 442
    - Alcelaphinae, 425–426
    - Antilopinae, 421–422, 429
    - Bovinae, 420–421, 429
    - Caprinae, 425–425, 429
    - Cephalophinae, 422–423
    - Hippotraginae, 425–426
    - Reduncinae, 425–426
  - Cervidae, 426–428, 429
  - Moschidae, 426–428
  - Hippopotamidae, 269, 412, 428
  - Suidae / Suiformes, 428–429, 442
  - Tragulidae, 428–429

**Mammalia (continued)**

Carnivora, 3, 5, 29, 273, 283, 366, 370, 411, 413, 415, 442, 449, 452, 454, 474  
*Ailurus fulgens*, 273  
Felidae, 335  
Herpestidae, 455  
*Odobenus rosmarus*, 273  
Cetacea, 11, 269, 411–412, 414, 429, 520  
Balaenopteridae, 442  
Delphinidae, 442  
Cetartiodactyla, 371, 379, 381  
Chiroptera, 29, 366, 370–372, 413, 440, 442, 451  
*Craseonycteris thonglongyai*, 451  
Kerivoulinae, 441  
Miniopterinae, 441  
Murininae, 441  
*Myotis*, 451  
Natalidae, 441  
Phyllostomidae, 440  
Rhinopomatidae, 441

**Mammalia (continued)**

Equidae, 442  
Leporidae, 442  
Marsupialia, 96, 373, 442  
Primates, 5, 9, 11, 29, 254, 256, 273, 276, 281–282, 288–294, 366, 370, 382, 411, 413, 415, 442, 446–449, 452, 474, 487, 489, 502, 514, 517–522, 525  
Cercopithecinae, 291, 448, 503–504, 514, 520, 522  
Colobinae, 291, 514  
Hominidae, 254, 256, 282, 503, 514, 516, 520  
Platyrrhini, 514, 516–517  
*Ramapithecus*, 282  
Strepsirrhini, 514–515, 520  
Rodentia, 259, 373  
Caviidae, 442  
Muridae, 259, 442  
*Schistosoma*, 29  
Vertebrata, 249, 256

# Subject index

*ad hoc*, 23, 107, 122, 353, 355, 358, 360, 381  
adaptation, 11, 411, 439–440, 449, 463, 466–467, 471, 477–478  
algorithm  
  ANCESTRALBUILD, 165–170  
  bi clique enumeration, 79, 258  
  branch-and-bound, 22, 48  
  BUILD (*also ONE TREE*), 7–8, 67, 70, 129–130, 132, 137, 142, 144–145, 147–148, 152, 154, 230, 248, 260  
  clique-finding, 48, 60  
  DESCENDANT, 166–168  
  DYADIC TREE CONSTRUCTION, 349  
  exact, 68, 76, 129  
  FCC-approximation, 67–68  
  FULLY-LABELEDBUILD, 157–159, 162–164, 168  
  heuristic, 76  
  large-tree, 494  
  least-squares, 90, 95, 98  
  maximum parsimony, 66, 92, 217, 380  
  MCMC, 193, 222  
  Metropolis, 218  
  MINCUTSUPERTREE, 67, 76, 80, 108, 221, 240, 247, 249, 260–262  
  modified MinCut, 67, 76, 80, 262  
  MRP, 19, 248  
  RANKEDTREE, 7, 131–138, 140–142, 148  
  reconciled-tree, 110

algorithm (*continued*)  
  SEMI-LABELEDBUILD, 153–155, 157–164, 166, 170  
  size-sensitive ERM taxon-addition, 514  
  small-tree, 494  
  strict-consensus merger, 311  
  strict-supertree, 22  
  sum-of-absolute differences, 91  
  supertree, 108, 151, 153, 248, 272  
  taxon-addition, 523  
  TBR, 44  
  tree-building, 248, 295  
  UPGMA, 372  
  allozymes, 412, 414, 425–426  
  amino acid, 18, 78, 293, 378  
  ancestor, 39, 154, 165, 445, 499  
    common, 160, 444, 518  
    immediate, 252  
    most recent common (*also lowest common*), 73, 131, 154–155, 164–165, 260  
    proper, 137, 169–170  
  APG (Angiosperm Phylogeny Group), 474–475  
  axiom, 8, 100, 117, 228, 231–233, 236–238  
  co-Pareto, 100, 117, 237–238  
  Pareto, 117, 236  
  basal, 21, 40, 59, 235, 259, 363, 396–397, 404, 428–429, 504, 507

- bias  
 phylogenetic, 519–520  
 positional, 119, 235, 363  
 region-specific, 18  
 size, 57, 118–119, 233–234, 353, 357, 364  
 tree-shape, 234, 363  
 trickle-down, 507, 512, 516, 522
- bijection, 152, 333
- biodiversity, 18, 451, 453, 455, 466
- biogeography, 7, 109–110, 204, 524
- bipartition  
 factor, 193, 199–218, 220–222  
 frequency, 193, 198–202, 206, 212, 214, 216, 220, 222  
 odds, 199, 210, 212, 214–215  
 taxon, 8, 193, 198–206, 209–210, 213, 215, 217, 219–220, 310, 343
- black box, 375–377
- branch (*see also edge*)  
 central, 202  
 internal, 230, 355, 504, 507–514, 523  
 length, 7, 76, 87–89, 92, 96–101, 196, 202, 219, 238–239, 281–282, 284–285, 289, 374, 419, 443, 447–448, 451, 453–454, 488–489  
 peripheral, 202
- branch swapping, 187, 338, 380  
 NNI, 39–40, 59, 120–1213  
 TBR, 78, 180–182, 381, 418
- branching event, 283, 285, 488, 490, 499, 510
- branching process, 490, 499, 504–505, 510
- Cambrian explosion, 283
- cardinality, 79, 237, 341, 343–344, 348–349
- character (*see also trait*)  
 ancestral, 273, 489  
 binary, 10, 22, 25, 37, 40–42, 51–52, 55, 58, 142, 145–148, 203–205, 208, 214, 217, 220–221, 230, 309, 310
- character (*continued*)  
 binary (*continued*), 354, 364, 389, 392–393, 401–402, 404  
 cladistically informative, 355, 379  
 compatible, 35, 37, 40–41, 43, 51, 55, 57, 395  
 continuous, 447, 449  
 derived, 25, 146, 359  
 discrete, 18, 22, 238, 248, 374, 377, 380, 449  
 incompatible, 50–51, 57, 72, 74  
 key, 464  
 matrix element, 24, 70, 108, 230, 272, 354–364, 374, 376, 379, 384  
 partial binary, 309–310  
 primitive, 25, 146–147, 359  
 pseudocharacter, 230, 233, 238–239, 241  
 shared derived, 283, 353, 355–360, 374, 392–393, 404, 468  
 shared primitive, 355  
 state, 37, 40–42, 66, 145, 241, 268, 288, 291, 355, 359, 374, 378, 449, 468, 475–476, 479, 481
- clade  
 age of, 282–283  
 basal, 40, 59  
 inclusive, 66, 502, 522  
 ingroup, 504–505, 508–509, 512–513  
 membership, 37, 66, 283, 359  
 nested, 502  
 novel, 337, 361–363, 365, 376, 384, 413  
 outgroup, 504, 505, 507–509  
 parent, 282  
 polymorphic, 471  
 resolved, 98  
 sister, 25, 40, 80, 95, 116, 118, 230, 241, 275, 282–283, 412, 421–429, 443–449, 468, 471–478, 480, 490–491, 505, 507, 518, 524  
 size, 66, 443, 448, 468

- clade (*continued*)  
 subclade, 504, 507  
 unresolved, 80
- cladistics, 10, 61, 276, 353–366, 441
- cladogenesis, 43, 283, 447–448, 452, 480, 492, 494, 497, 499, 525–526
- classification, 101, 108, 130–131, 202, 247, 251–254, 256, 282–283, 377–379, 382, 452, 461–462, 476
- clique, 40–44, 47–49, 230, 307–308, 364, 396  
 best, 41–42, 47, 49–50, 57–59  
 biclique, 9, 78–82, 258–259  
 maximal, 79–80, 230, 258  
 maximal, 307  
 maximum, 35, 41–43, 47–52, 57–60  
 ultraclique, 59–61
- cluster  
 maximal, 160, 162–163  
 minimal, 160, 163  
 proper, 68–69, 260–261  
 subcluster, 130, 158
- coding  
 additive binary, 4, 10, 19, 25–26, 30, 35–37, 57, 65, 108–109, 118, 204, 230, 310, 354, 390, 394, 468, 475  
 Purvis, 25, 119, 233  
 quartet, 347–348  
 three-item, 394
- co-evolution (*also co-phylogeny; co-speciation*), 7, 109–111, 204, 411, 440, 442–443, 449–450, 467, 478, 525
- combinatorics, 346
- comparative method, 96, 229, 231, 281, 369, 382, 439–444, 447, 453–454, 463  
 independent contrasts, 468, 475–476, 479–480, 510
- compatibility  
 ancestral, 165–169  
 collective, 40–41  
 fractional character, 67  
 joint, 22
- compatibility (*continued*)  
 of characters, 35, 37, 40–43, 50–51, 55, 57, 59, 67, 70–74, 142, 145–148, 230  
 of data, 18, 24, 194, 365–366, 374, 412  
 of methods, 28, 242, 375, 384  
 of partitions, 35, 55, 66, 73, 200, 202, 206, 213–214, 237, 370, 379, 381, 462  
 of quartets, 144–145  
 of results, 313  
 of taxa, 153–154, 166, 170  
 of trees, 4, 19, 22, 24, 27–30, 47, 65–67, 69–70, 72, 74, 96, 98, 108, 115, 122, 134–136, 143–144, 147–148, 152–154, 158–159, 161–165, 167–169, 230, 311, 359, 369, 375–376, 381, 461  
 pairwise, 40
- completion, 70  
 chain graph, 72
- component, 117, 136, 158, 230, 233, 236, 238, 257, 259–260, 347, 354–356, 389–404  
 arc, 167–169  
 connected, 68, 79, 89, 158, 160, 163–164, 167, 255, 257–258, 260–261, 306, 347  
 partial, 400
- computational complexity (*see also running time; speed*), 23, 70, 73, 111, 136, 138, 144–145, 178, 180, 186, 215, 218, 220–221, 240, 332, 342, 344, 349  
 exponential-time, 22, 60, 68, 240  
 fixed-parameter tractable, 68, 73, 112  
 linear-time, 52, 112, 186, 218  
 NP-complete, 67–68, 71–72, 108, 112, 143, 145, 170, 176, 248–249, 332, 342–343, 349

- computational complexity (*continued*)  
NP-hard, 35, 48, 100, 142, 306–307,  
312  
polynomial-time, 22, 52, 60, 67, 70,  
72, 112, 117, 130–133, 136, 138,  
142–144, 146–148, 151–154,  
164–165, 170, 194, 221, 240,  
248–249, 307, 338, 342–343, 349  
quadratic, 333  
running time, 136, 138, 142, 145, 164,  
170, 178, 180, 186, 218, 301,  
303–304, 315–316, 318, 333, 338,  
343–344, 349  
speed, 39, 48, 52, 60, 117, 179,  
186–187, 221, 227, 235, 240, 302,  
306, 324  
computer cluster, 315  
computer programs  
3item, 396  
ape, 289  
BioPerl, 40  
CLINCH, 40, 48  
Clustal W, 288  
Component, 430  
formatdb, 288  
LEDA, 314  
ModelTEST, 288, 314  
MrBayes, 22, 209  
PAUP\*, 44–46, 76, 78, 98, 113, 178,  
180–182, 187, 206, 260, 289, 303,  
308–309, 312, 315, 320, 381, 415,  
418, 468  
PAUPRat, 288  
Perl, 40, 48–49, 273, 289, 314–315,  
418  
perlRat, 309, 418  
PHYLIP, 44, 48  
  CLIQUE, 44, 48  
  MIX, 44, 46  
Quartet Suite, 188  
R, 289, 470  
r8s, 30, 74, 180  
computer programs (*continued*)  
RadCon, 30, 174, 288  
SplitsTree, 55  
Supertree, 30, 113  
SuperTree, 30, 468, 475  
SYMMETREE, 487, 490, 502, 514,  
522–523  
synonoTree.pl, 273  
TAX, 396  
congruence  
character (*see also supermatrix; total  
evidence*), 29, 354, 362, 401, 404  
global, 92  
taxonomic, 355, 401, 404  
consensus  
Adams, 74, 113, 230, 241, 249, 390,  
403  
average, 6, 87–101, 229, 375  
majority-rule, 73–74, 87, 92, 182–183,  
228, 231, 234, 236, 288, 291, 390,  
419  
median, 90  
Nelson, 390  
network, 91  
reduced, 236, 403  
semi-strict, 59, 73–74, 390  
spectral, 91  
strict, 4, 42, 73–74, 78, 87, 98, 113,  
180, 228, 272, 288, 310–311, 337,  
340–341, 362, 390, 397–398,  
401–402, 419, 421, 426  
supertree, 42–43, 184, 236, 249, 289,  
376, 419, 421, 426  
three-item, 402  
consensus setting, 27, 87–88, 91, 93, 95,  
101, 228, 389  
conservation, 281, 411–412, 414, 421,  
451, 453, 455  
consistency, 40–42, 66, 281  
  local, 173–174, 177–178, 187  
measure, 25, 30  
of characters, 22, 51, 60, 380

- consistency (*continued*)  
 of classifications, 247, 253, 256, 262  
 of data, 29, 389, 401, 403  
 of partitions, 41, 65  
 of quartets, 173–174, 176–177, 187  
 of rooting, 25  
 of trees, 28, 35, 37–47, 50, 55–59, 65,  
   96, 198, 201–202, 212–213, 215,  
   252–253, 256, 260, 369  
 taxonomic, 250–251, 253, 273
- consistency index (CI), 291, 353, 356,  
 360–362, 364
- constraints (*see also search, constrained*)  
 computational, 39, 52, 55, 333, 480  
 precedence, 133–134, 136, 139  
 ultrametric, 91
- date  
 age, 282–283, 444–445, 447–448, 478,  
   490, 519, 520, 525  
 minimum, 282  
 divergence time, 9, 11, 129, 131–132,  
   137–142, 148, 281–285, 287, 289,  
   291–294, 443–444, 446, 448, 451,  
   453, 487–488, 516–521, 524  
 in Primates, 293  
 relative, 129, 131–133, 137–138,  
   140, 148, 281, 288  
 fossil, 282–283, 453, 467, 519–520
- decomposition, 9, 94, 186, 188, 301–325  
 disk-covering method, 303–325  
 quartet, 176, 186, 304  
 random, 303, 308, 312–313, 315,  
   317–319, 325  
 split, 55, 61, 91  
 triplet, 186
- degree, 68, 71, 130, 143–144, 147, 152,  
   161, 164, 175, 305, 310, 333–334  
 in-degree, 167–169  
 maximum, 72  
 out-degree, 152, 161, 253–254
- descendant, 130, 154  
 proper, 73, 131, 160
- disk-covering methods, 302–325, 332
- displays, 68–69, 73–74, 117, 130,  
   132–134, 138, 143–148, 152–154,  
   165, 175–176, 179, 187, 237, 259,  
   334, 338, 345, 347  
 ancestrally, 154, 165–169  
 perfectly, 153–154, 158–165, 170
- distribution, 193, 220, 490  
 binomial, 120  
 clade size, 468  
 diversification, 77, 521, 524, 526  
 exponential, 286  
 geographic range, 446, 455  
 joint-probability, 219  
 negative exponential, 285  
 non-random, 452  
 non-uniform, 471  
 partition metric, 44, 46  
 posterior-probability, 121, 195–196,  
   198, 218–219, 523  
 prior-probability, 121, 194, 196, 199,  
   216–217, 220  
 probability, 51, 121, 193, 195–199,  
   211, 216, 219–220, 285, 468  
 proposal, 218  
 quartet-fit similarity, 182–183  
 rate, 499  
 taxon, 78  
 temporal, 488  
 topological, 488, 491, 494  
 tree scores, 221  
 uniform, 490, 501, 526
- divide-and-conquer, 9, 248, 301–305,  
   308, 312–313, 316, 320, 324–325
- DNA, 9, 18, 27, 58, 66, 68, 74, 97–98,  
   121, 180, 248, 250, 268, 281, 283,  
   288, 304, 314, 374, 378, 383, 462,  
   472, 489, 519, 525  
 mitochondrial, 271, 292  
 ribosomal, 314, 413–414, 425–426,

- DNA-DNA hybridization, 194, 248, 365, 372, 374, 377–379  
 dyadic closure, 187, 332, 348–349  
 edge (*see also branch*)  
   arc (directed edge), 145–146, 166–167, 169, 255  
   incident, 167–168  
 blue, 155–157  
 chord, 306  
 incident, 168  
 interior, 68, 144, 333–334, 337, 341–342, 347–348  
 red, 156–157, 164  
 Edgington's combined probability test, 493–494  
 error, 65–67, 76–77, 79, 82, 112, 121–122, 188, 251, 282, 440, 450, 453–454, 475, 480, 514, 519–521, 523  
 estimation, 95, 112–113, 122, 521  
 sampling, 37, 453  
 Type I, 288, 468, 471, 500, 510–512, 514  
 Type II, 468  
 excess, 177, 342–343, 345  
 excess-free, 332, 341–349  
 extinction, 96, 110, 281, 372, 411, 414, 417, 440, 442, 445–446, 451–454, 464–466, 468, 480, 488, 490, 523  
 Fisher's combined probability test, 468, 471, 492–494  
 GenBank, 78, 251, 269, 288  
 gene  
   12S rDNA, 371, 413–414  
   16S rDNA, 314, 413–414  
   18S rDNA, 314, 475  
   28S rDNA, 114, 116–117  
   *Alcohol dehydrogenase*, 114–115, 206  
   *Alcohol dehydrogenase related*, 114  
   *atpB*, 314, 475  
   *Cu-Zn superoxide dismutase*, 114  
   cytochrome *b*, 413–414, 425–426  
   cytochrome *c*, 414
- gene (*continued*)  
   *Dopa decarboxylase*, 114, 117  
   in Artiodactyla, 416–417  
   in Primates, 290–291  
   *rbcL*, 304, 314, 320–321, 323, 464, 475  
   *roughex*, 114  
 gene duplication, 110–112, 114–115, 117, 119, 121–122, 376  
 gene loss, 110–111, 114, 117, 119, 121–122, 376  
 gene transfer, 24, 30, 110–111, 121, 376  
 generation time, 464, 472  
 geography, 401, 445–447, 453, 455  
 graph (*see also tree*)  
   Σ-subgraph, 71–72  
   acyclic, 253  
   bipartite, 71–72, 78–79, 258  
   cladogram, 353–354, 372, 389–404  
   classification, 8, 247, 252–254, 256, 273  
   cluster, 257, 259  
   cluster and root-label, 155, 157–160, 162, 164, 166  
   connected, 136, 253, 257, 260, 307–308, 333, 348, 404  
   dendrogram, 88, 91, 94  
   descendancy, 166–167, 169  
   directed, 252, 255  
   phenogram, 415  
   split, 35, 55, 61, 91  
   subgraph, 79, 134, 167, 254–256, 258, 307, 333, 348  
   threshold, 306, 308, 314–315  
   triangulated, 306–308  
 graph theory, 65, 68, 71–72, 77, 331, 376  
 Hennig's dilemma, 395, 403  
 hidden support (*see also signal enhancement*), 375, 377, 384  
 hierarchy  
   amalgamation, 344, 346  
   maximal, 346

- hierarchy (*continued*)**  
 ranked, 133–136  
**homology**, 361, 377–378, 392–394  
 homologue, 358, 393  
 orthology, 28, 78, 378–379  
**homoplasy**, 35–37, 40, 66–67, 145, 147, 205, 356, 358–361, 376, 380, 412, 451  
 convergence, 37, 359–360, 376, 442, 468  
 parallelism, 37, 66, 468  
 paralogy, 28, 117, 378, 404  
 reversal, 37, 66, 359–360, 376, 397, 468  
 horizontal transfer, 110–111, 121, 450  
 hybridization, 28, 110, 465  
 immunology, 27, 371, 374, 377–379  
 incongruence, 23–24, 58–59, 98, 113, 115, 117, 122, 216, 354, 358–362, 376, 413  
 incongruence length difference test, 98  
 independence, 198, 232–233, 270, 281–282, 359, 363, 401, 499  
 of irrelevant alternatives, 117, 232  
 of models, 219–220  
 of nodes, 21, 109, 220, 239, 273, 353, 364, 442, 449, 463, 522  
 of trees, 21, 57, 92, 96, 217–218, 249–250, 267–272, 274, 353, 355, 363–364, 370–371, 376, 378, 413  
 size, 57, 218  
 statistical, 442, 449, 470, 493  
**key innovation**, 11, 448, 461, 463, 471, 475–477, 480, 504  
**label**, 65–66, 153–164, 167–170, 251–252, 260, 332, 334–335, 339  
 ancestor, 155  
 descendent, 155, 162–163  
 interior, 151, 251–252, 257, 273  
 root, 154, 156, 160, 162–163, 168–169  
 terminal, 256  
**labeling**, 8, 19, 113, 151–154, 156–157, 160–164, 167–168, 170, 251, 262, 289, 344  
**long-branch attraction**, 373, 429, 443  
**macroevolution**, 5, 11, 439, 461–463  
**Markov chain Monte Carlo (MCMC)**, 22, 121, 188, 193, 195, 198–199, 202, 209–211, 214–218, 220–222  
**matrix**  
 character, 5, 18, 23, 173–174, 194, 249–250, 356–357, 360–362, 364–365, 371–372, 379–380, 390, 401–402, 415  
 distance, 89–90, 92, 95–96, 98, 314, 374  
 additive, 94  
 average, 90, 93, 95, 98  
 branch, 99  
 incomplete, 94–96  
 median, 91  
 path-length, 6, 88–89, 91, 94–95, 98–100, 229, 241  
 ultrametric, 241  
 three-item, 396  
 ultrametric, 94  
**maximum likelihood**, 96, 120–121, 174, 188, 194, 221, 292, 301–303, 305, 313, 325, 373–375, 390, 412, 504, 510, 520  
**meta-analysis**, 194, 229, 462  
**mini-supertree**, 272, 275  
**model**  
 age and area, 445  
 among site variation, 77, 209, 314  
 character, 442–444, 454, 489  
     Brownian motion, 444  
     punctuational, 444  
 character-state change, 288  
 cladogenesis, 480, 492, 499, 501, 504, 521, 525  
 deterministic branching clock, 521

- model (*continued*)**
- cladogenesis (*continued*)**
    - equal-rates Markov, 74, 77, 180, 282, 285, 295, 448, 468, 474, 480, 490–491, 494, 497, 499, 504–505, 509–510, 514, 517
    - gradual, 499, 502, 526
    - one parameter, 504–509
    - punctuational, 499, 501–502, 526
    - two parameter, 504–509
  - coalescent, 292
  - evolutionary (*see also rate, evolution, molecular*), 18, 22–24, 77, 112–113, 121, 218, 221, 281, 289, 292, 294–295, 314, 383, 412, 447
  - gene duplication / gene loss, 121
  - general time reversible, 209, 219
  - Jukes-Cantor, 98
  - Kimura two-parameter, 180, 314
  - maximum likelihood, 412, 504
  - NNI-binomial, 120–121
  - null, 121, 444, 446, 448, 474, 480, 491
  - phylogenetic, 196, 219
  - speciation, 445
  - stochastic branching, 490
  - substitution, 196, 219, 288–289, 295, 314, 489
  - molecular clock, 281, 283–284, 288–289, 294–295, 491, 521
  - monophyly, 80, 115, 273–275, 277, 312, 356, 380–381, 383, 411, 414, 417–418, 420–421, 427–428
  - neighbour joining, 304–306
  - nesting, 73–74, 230, 236, 261
  - numerical taxonomy, 392
  - optimization
    - ACCTRAN, 468, 470–471
    - DELTRAN, 468, 470–471
  - order, 69
    - filling, 95
    - genome, 248
  - order (*continued*)
    - hierarchical, 19, 21, 58, 130–132, 139–140, 252–254, 256, 283, 287
    - input, 117, 235, 239, 308–310, 312, 335–336, 344, 454
    - linear, 255–256
    - partial, 130–131, 154–166
    - taxon, 44
  - outgroup, 79, 373, 504–505, 507–509
    - all-zero, 19–21, 354, 412
  - overlap
    - data, 78, 268, 270–271, 363, 370–371, 373, 378
    - geographic, 445–446, 453
    - joint, 77, 79
    - minimal, 257, 259
    - taxon, 257
    - taxonomic, 3–4, 9, 17, 27–28, 35, 42, 54–56, 58–60, 75, 78–80, 82, 88, 93, 97–98, 107, 129, 151, 173–174, 180–182, 188, 194, 217, 229, 234, 236, 247, 251, 253–254, 257, 259, 262, 267, 271, 302–306, 308, 318, 320, 323, 332, 354, 365, 389, 415, 421, 461
  - paraphyly, 115–116, 269, 373, 412, 426, 475
  - parsimony
    - Dollo, 25, 69
    - Fitch, 230
    - Wagner, 69, 230, 395
  - parsimony ratchet, 288, 301, 303–304, 309, 320–324, 418
  - patch, 348
  - patchwork, 333, 345–346, 348
    - ample, 346–347
  - phenotype, 18, 446–447, 451, 463, 471–472, 479
  - phylogenetic diversity (*PD*), 96, 451–453
  - polymorphism, 468, 471, 475–476, 478, 525
  - polyphyly, 115, 421, 475

- polytomy, 21, 23, 43, 55, 118, 130, 152, 154, 164–165, 230, 312, 426, 429, 514, 516, 522–523  
 hard, 43  
 soft, 108, 523  
   collapsed, 523  
   consensus, 523
- power, 11, 53, 59–60, 210, 272, 295, 395, 454, 463, 479, 487–489, 491–492, 495–497, 499–503, 510, 512–514, 524  
 resolving, 202
- probability  
   branch length, 219  
   branching, 505  
   conditional, 21  
   cumulative, 468, 490–491, 494–495, 509  
   deletion, 75–77, 180, 183  
   density, 198  
   diversification, 468, 490, 499  
   rate shift, 504, 509, 516  
   equal-rates Markov, 494  
   equal-rates Markov nodal, 491–498, 510, 512  
   extinction, 465  
   joint, 219  
   marginal, 195, 216, 219  
   NNI, 120–121  
   non-informative, 219  
   point, 494  
   posterior, 121, 188, 195–196, 198–199, 202, 210, 218–220, 523  
   maximum, 202  
   prior, 121, 188, 193–196, 198–199, 210, 216–217, 219–220  
   uniform, 196  
   proposal, 210  
   ratio, 199  
   relative, 196, 198, 200, 211, 215, 218, 220  
   space, 494, 506  
   speciation, 477
- probability (*continued*)  
   topology, 219, 509  
   T-PTP, 374  
   tree, 120–121, 193, 196–199, 206, 209–220, 494–495, 497
- problem  
   black-box, 375  
   chain graph completion, 72  
   classification graphs, 256  
   cluster graphs for higher taxa, 257  
   compatibility of binary characters, 145  
   compatibility of unrooted trees, 143  
   consensus tree, 78, 395  
   consistency of trees with labeled internal nodes, 256, 262  
   discreteness, 494  
   edge modification, 72  
   excess-free subset, 332, 349  
   filling order, 95  
   fractional character compatibility, 67  
   gene tree-species tree, 110  
   HIGHER TAXA ANCESTOR COMPATIBILITY, 154, 166  
   HIGHER TAXA COMPATIBILITY, 153–154, 170  
   limitations of MINCUTSUPERTREE, 262  
   maximum edge biclique, 79  
   maximum likelihood, 301–302  
   maximum parsimony, 108, 301–302  
   maximum quartet consistency, 176  
   maximum-clique, 48, 60  
   minimization, 72  
   missing distances, 95  
   MRF, 66–72  
     decision, 71–73  
     weighted, 70  
   MRP, 66  
   non-independence, 371, 373, 494  
   optimal tree refinement, 312  
   parent tree, 332, 342–343, 349  
   PHYLOGENETIC DIVERGENCE TIMES, 138, 140, 142

problem (*continued*)

- PHYLOGENETIC RANKING, 131
- quality control, 372
- quartet compatibility, 144
- sandwich-to-ultrametric, 138
- supertree, 28, 68–69, 78, 108, 119, 121, 142, 146, 249, 331–332, 395, 397–398, 404
- rooted, 331
- unrooted, 331, 335, 349
- tree-building, 66
- trickle-down, 507, 512, 516, 522
- unique identifiers for taxa, 251, 262
- protein, 18–19, 23–24, 27, 78, 206, 248, 378, 412, 520
- quartet, 8, 22, 94, 144, 173–181, 185–188, 230–231, 236, 302, 304–305, 325, 334–341, 343–344, 347–349
- missing, 174, 176, 178, 186–187, 231
- number of, 178, 186
- quartet cleaning, 187, 304
- quartet puzzling, 96, 144, 177, 187, 228, 231, 236, 239, 302, 304
- quintet, 179
- radiation (*also adaptive radiation*), 472, 474, 477–478, 516, 520, 522, 525
- rate
  - branching, 505, 510
  - cladogenesis, 448, 463, 477, 525
  - diversification, 11, 281, 448, 463, 465, 471, 474, 477–481, 487–526
  - parameter, 504–507
  - error, 500–512, 514
  - evolution, 37, 47, 58, 97–99, 372, 383, 442, 447, 449, 454, 463–465, 471, 489
  - gene, 111
  - molecular, 24, 77, 209, 219, 283, 288–289, 464–465, 472, 489, 491, 492, 525
  - morphological, 489

rate (*continued*)

- extinction, 480, 523
- homogeneity, 491
- reproduction, 464
- speciation, 449, 461, 464–466, 468, 470–472, 474, 480, 523
- refinement, 130, 134, 137, 143, 152, 165, 170, 302–303, 305, 332, 338
- optimal tree, 312–325
- proper, 160
- RNA, 27, 304, 314, 320, 322–323, 371
- root, 39–41, 68, 88, 143–147, 152, 154–155, 158, 166–170, 181, 230–231, 241, 253–254, 260, 283, 285–286, 289, 291, 331, 336, 373, 468, 497, 505, 507, 514–516, 518
- pseudo, 337
- sampling
  - biclique, 79
  - Gibbs, 218
  - MCMC, 198–199, 202, 209, 211, 218, 221–222
  - of characters, 239, 374, 380
  - of posterior probabilities, 195, 198
  - of trees, 77–78, 114, 121, 188, 193, 202, 211, 214–215, 239, 474
- taxon, 47, 76–77, 79–80, 82, 96, 185, 250, 262, 303, 366, 440, 447–448, 451, 462–463, 472, 474–475, 480, 487–488
- search
  - BLAST, 78, 288
  - bootstrap, 206
  - branch-and-bound, 48
  - constrained, 206, 210, 215, 247, 259–261, 289, 312, 315, 418
  - constraints, 259
  - gene tree parsimony, 113, 114
  - global, 323
  - grep, 288
  - hill-climbing, 187
  - iterative, 48

- search (*continued*)
- maximum likelihood, 305
  - maximum parsimony, 44, 76, 81, 113–114, 178, 180–183, 187, 303–306, 308–310, 312, 315, 317, 319–320, 324, 380–382
  - MCMC, 22
  - parsimony ratchet, 288, 309, 320, 418
- set
- arc, 166
  - bipartition, 310
  - bootstrap, 249
  - branch, 516
  - branch length, 219, 489
  - character, 35, 37, 40–41, 43, 55, 59, 188, 271, 310, 355, 364, 370, 372, 395
  - cluster, 65, 130, 132–133, 155
  - connected, 38
  - constraint, 133
  - cut, 260
  - distance, 95
  - edge, 71, 134, 155, 162–163, 166, 260–261, 348
  - event, 111
  - factor, 193
  - flip, 67
  - label, 154–156, 159–163, 170
  - lineage, 517, 519
  - nested, 288
  - node, 78, 258, 285–287, 291
  - numeric, 131, 133, 137, 139, 140, 147, 339, 345
  - parameter, 39, 44, 46, 49, 51, 53, 55–56, 59, 74, 97, 180, 182, 194–196, 219–220, 308, 312–314, 318, 499, 504, 510, 521, 523
  - partition, 67
  - patch, 348
  - probability, 493
  - quartet, 175–176, 187, 231
  - rejected, 178–179
- set (*continued*)
- quartet (*continued*)
    - supported, 178–179
  - relationship, 270
  - superset, 271
  - time, 131, 138
  - triplet, 134
  - vertex, 71, 131, 134, 137, 139, 147, 152, 154–155, 157–158, 160, 162–163, 166–168
  - weight, 95
- signal, 178, 283, 361, 375–376, 443–444, 475
- conflicting, 43, 118
  - primary, 360–361, 365
  - subsignal, 360–362, 365
- signal enhancement (*see also hidden support*), 249, 271
- simulation
- Monte Carlo, 37, 188, 494, 500, 512
  - sorting, 28, 255, 378, 393
    - lineage, 28
    - topological, 255
  - speciation, 109–110, 131–132, 137–138, 281, 440, 444–446, 449, 461, 463–468, 470–472, 474, 477–478, 480, 490, 499, 501, 523
  - allopatric, 446
  - sympatric, 446
  - species concepts, 455
  - species richness, 11, 448–449, 453, 461, 463–468, 471–472, 474–476, 478–481
- supermatrix (*see also congruence, character; total evidence*), 5, 10, 193, 221–222, 369, 372–373, 375, 377–384, 412, 429
- supertree methods
- average consensus, 6, 87–101, 229–230, 238–239, 241, 375
  - Bayesian, 193–222
  - WAB, 193, 205–206, 208–211, 214–218, 220–222

supertree methods (*continued*)Bayesian (*continued*)

WIB, 193, 199, 202, 204, 206,  
208–214, 217, 220

gene tree parsimony, 6–7, 107, 109,  
111–119, 248–249

MINCUTSUPERTREE, 6, 7, 27, 65,  
67–68, 74–77, 79–80, 82, 108, 118,  
148, 170, 221, 230, 233, 235–236,  
240–241, 247–249, 260–262, 375

Modified MinCut, 6–7, 65, 67–68, 74,  
76–80, 82, 118, 262

MRC, 6, 20, 35–61, 69, 74, 230,  
236–237

triplet, 231, 240

MRF, 6, 9, 65–82, 122, 230, 234, 236,  
248, 262

MRP, 4–6, 8, 10, 17–30, 35–37, 40,  
42–44, 46–47, 51, 53–61, 65–69,  
74–77, 79, 81–82, 91, 96, 99, 101,  
108–109, 113–119, 129, 146,  
173–174, 178, 180–185, 187, 194,  
204, 206, 217–218, 220–221, 230,  
233–237, 239, 241, 248, 260, 262,  
268, 270, 272, 281–282, 284, 288,  
295, 301–304, 308–310, 312–313,  
315, 317–319, 325, 353–366,  
369–384, 390–391, 395–396,  
402–403, 412–413, 415, 418,  
429–430, 462, 468, 480

irreversible, 26, 69, 146, 230,  
234–235, 237, 241, 291,  
359–360, 468, 475

Purvis coding, 25, 119, 230,  
233–235, 237, 239, 241

quartet, 174, 230–231, 241

triplet, 230–231, 235, 240

weighted, 25, 113, 218, 221, 238,  
358, 364, 468, 475

quartet, 8, 22, 144, 173–188, 231, 235,  
241

QILI, 173–174, 176, 180–188

supertree methods (*continued*)quartet (*continued*)

QLI, 173–174, 176, 180–182,  
186–188

RANKEDTREE, 7, 131–138, 140–142,  
148

semi-strict, 6, 229, 233, 237, 369–370,  
375, 377, 381, 384

strict, 6–7, 22, 229, 233, 311,  
369–370, 377, 381, 384, 389, 393

strict consensus merger, 301, 303, 304,  
307, 310–313, 315–320, 323, 325

supertree setting, 6–7, 87–88, 92–93,  
96–97, 101, 121

## support techniques

bootstrap, 92, 113, 187–188, 205–206,  
214, 217, 221, 228, 238–239, 249,  
356, 363, 374–375, 468, 475, 523

Bremer's decay index, 193, 205–206,  
216, 221, 238–239, 356, 374,  
418–419, 424, 454

jackknife, 96, 205, 228, 239

## symbiosis, 28, 450

taxonomic coverage, 110, 270, 276,  
365–366, 384, 440, 462–463, 468

three-item parsimony, 398

three-item statement, 10, 389–404

time (*see date*)

total evidence (*see also congruence,*  
*character; supermatrix*), 5, 10, 18–19,  
23, 26–28, 88, 92–93, 100, 269, 271,  
354, 360–362, 364–365, 383,  
402–403, 413

trait, 30, 37, 96, 283, 372, 440, 442–444,  
447–449, 451, 454, 463–464,  
467–468, 470–472, 474–481  
life-history, 478

tree (*see also graph*)

asymmetric / unbalanced, 38, 46, 57,  
209, 234, 285, 287, 363, 452, 474,  
495

**tree (*continued*)**

- balanced / symmetric, 38–39, 44, 46, 57, 234, 363, 474, 495, 516
- binary (fully bifurcating), 21, 59, 144, 175, 177, 179, 198, 200–202, 206, 218, 233, 302, 305, 308, 312, 325, 332–334, 336–337, 339–340, 342–343, 345, 347, 523
- bootstrap, 113, 188, 205, 249, 523
- candidate, 310
- caterpillar, 343–344, 346–347
- character-state, 23, 26, 36, 364
- consensus, 43, 59, 73, 78, 87, 90–91, 101, 228, 232, 237, 239, 249, 341, 389, 391, 415
- Adams, 74, 113, 249
- average, 87–90, 92–93, 100
- majority-rule, 182, 288, 291, 421, 426, 429
- strict, 98, 113, 272, 288, 337, 340–341, 402, 421, 426
- constraint, 247, 259–260, 262, 289, 312
- dated, 11, 137–138, 148, 291, 451, 453
- fully-labeled, 155–156, 163, 166, 168–169
- gene, 23–24, 28, 37, 77–79, 82, 110, 112, 115–119, 248, 270, 376, 378, 404
- isomorphic, 153, 155, 333–335, 338–339
- MCMC, 188
- model / true, 35–61, 68, 74–80, 97–98, 180–183, 304, 307, 310, 313, 325, 413
- optimal
  - MAP, 202, 209
  - maximum likelihood, 373
  - maximum parsimony, 44–46, 57, 59, 66, 75–76, 78, 108, 114, 180, 182, 184, 291, 308, 310

**tree (*continued*)**

- maximum parsimony (*continued*)
  - 321–324, 355–356, 358, 373, 380–381, 402, 418
  - minimum flip distance, 76
  - quartet, 178
  - reconciled, 114
- parent, 108, 331–349
- pruned, 80, 232
- quartet, 144, 173–181, 186–188, 231, 305, 334–341, 343–344, 347–349
- quintet, 179
- random, 206, 210, 239
- ranked, 131–134, 137, 140–142
- reconciled, 109–114, 122
- reduced, 414
- reference, 202–206, 209–210, 215–217
- rooted, 19, 21, 25, 66, 68, 79, 88, 91, 108, 129–134, 137–148, 151–155, 180–181, 184, 186, 200, 231, 237, 241, 256, 260, 331, 334, 336, 341, 344, 346, 349
- fully-labeled, 155–169
- semi-labeled, 151–171
- scaled, 89
- semi-labeled, 154, 252
- skeleton, 188
- species, 18, 23, 37, 109–110, 112–113, 118–119, 270
- star (bush), 98, 181, 304, 313, 337, 397, 415
- starting, 38–39, 44–46, 76, 180, 210, 231, 239, 308–309, 320, 323, 418
- subtree
  - induced, 120, 332, 334, 349
  - maximum agreement, 75–77, 108, 236
  - minimal, 68, 130, 143, 147, 152, 175, 186
  - synthetic, 194, 372, 375, 381
- three-taxon, 41, 504–505, 508, 510

- tree (*continued*)  
 ultrametric, 88, 91, 283–285, 289  
 underlying, 154  
 unresolved, 181, 304, 312, 315, 332  
 unrooted, 10, 19, 28, 67, 91, 142–144,  
   175, 200, 210, 237, 241, 331–332,  
   335–336, 342, 355  
 weighted, 87–91, 93–97, 174, 221
- tree balance / shape, 21, 205, 234–235,  
   285, 362–363, 448, 452, 491–492,  
   495, 499, 502
- Tree of Life, 4–5, 18, 82, 151, 247–248,  
   262, 325, 384, 404, 452
- tree similarity measures  
   consensus fork index, 98–99  
   maximum agreement subtree, 75–77  
   partition metric, 44, 46, 202, 430  
   quartet-fit similarity, 75, 181–186  
   triplet-fit similarity, 75–76, 79–80,  
     181
- tree space, 22, 198–200, 202–206,  
   210–211, 218, 220–221, 288, 417–418  
   quartet, 175  
   supertree, 193, 218, 220, 222
- TreeBASE, 181, 247, 249–251, 269
- triplet, 75, 80, 94, 133–134, 144, 181,  
   186–187, 230, 233, 236–238, 240–241  
   number of, 186
- triplet puzzling, 231
- weighting, 17, 19, 24–27, 57–58, 95,  
   111–112, 117, 202, 205–206, 218,  
   233, 238–240, 260–261, 269, 306,  
   309, 357–358, 364, 373, 384, 418,  
   498, 510
- bootstrap, 205–206, 214, 221, 363,  
   374–375, 468, 475
- Bremer's decay-index, 206, 221, 374
- down, 272, 276
- inverse, 233
- maximal, 262
- of characters, 20, 24–25, 199, 205,  
   217–218, 221, 233, 238, 249, 353,  
   356, 362–364, 373–375, 384, 468,  
   475
- of data, 370, 413
- of evidence, 356–357
- of factors, 206, 210
- of flips, 70
- of nodal probabilities, 494–495, 498
- of quartets, 174, 176–178, 180,  
   186–188
- of trees, 19, 21, 25, 187, 233, 238,  
   249, 272, 276–277
- successive, 25