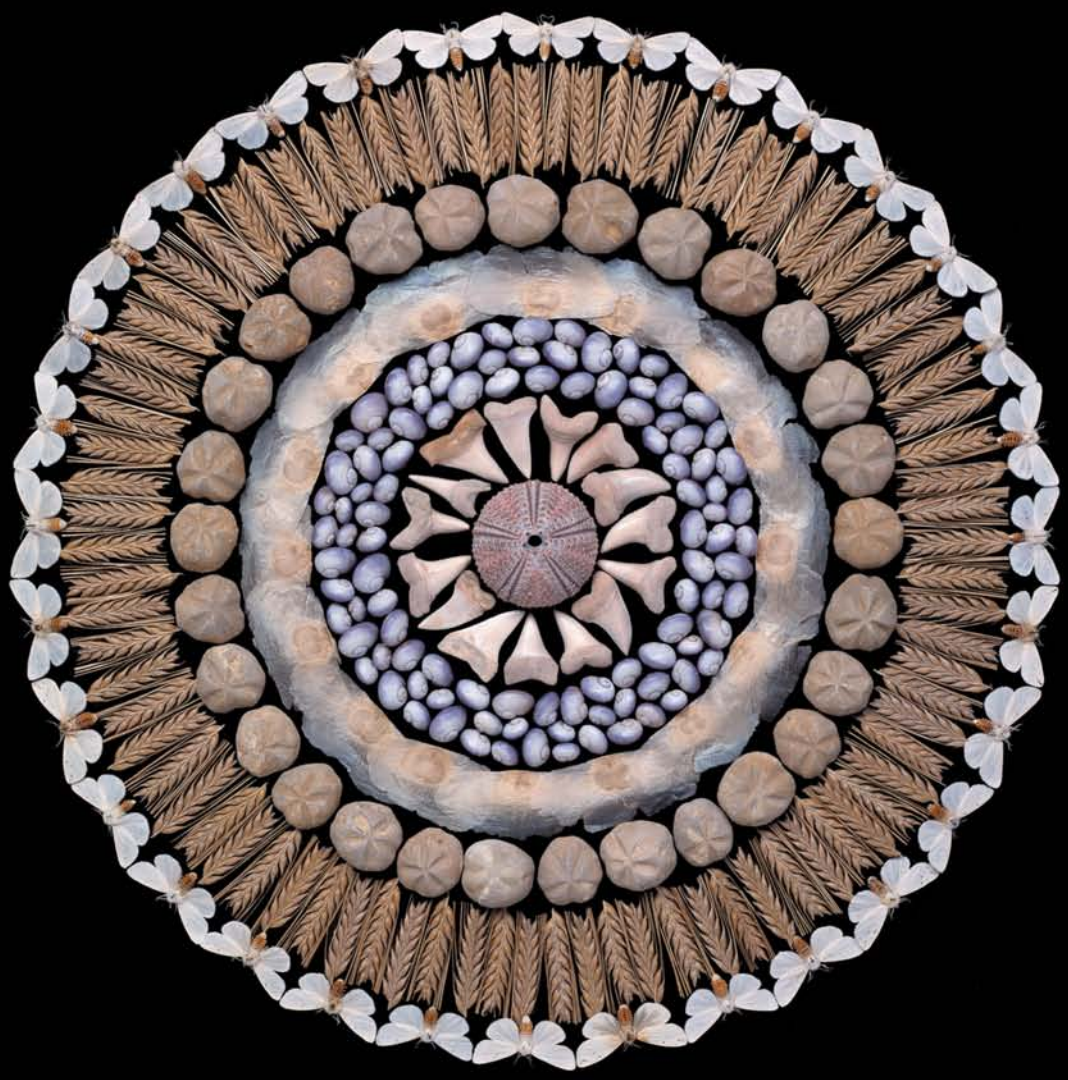
Reconstructing the Tree of Life

Taxonomy and Systematics of Species Rich Taxa



Edited by
Trevor R. Hodkinson
John A. N. Parnell

THE Systematics ASSOCIATION

CRC

The Systematics Association Special Volume Series 72

Reconstructing the Tree of Life

Taxonomy and Systematics of Species Rich Taxa

The Systematics Association Special Volume Series

Series Editor

Alan Warren
Department of Zoology, The Natural History Museum,
Cromwell Road, London SW7 5BD, UK.

The Systematics Association promotes all aspects of systematic biology by organizing conferences and workshops on key themes in systematics, publishing books and awarding modest grants in support of systematics research. Membership of the Association is open to internationally based professionals and amateurs with an interest in any branch of biology including palaeobiology. Members are entitled to attend conferences at discounted rates, to apply for grants and to receive the newsletters and mailed information; they also receive a generous discount on the purchase of all volumes produced by the Association.

The first of the Systematics Association's publications *The New Systematics* (1940) was a classic work edited by its then-president Sir Julian Huxley, that set out the problems facing general biologists in deciding which kinds of data would most effectively progress systematics. Since then, more than 70 volumes have been published, often in rapidly expanding areas of science where a modern synthesis is required.

The *modus operandi* of the Association is to encourage leading researchers to organize symposia that result in a multi-authored volume. In 1997 the Association organized the first of its international Biennial Conferences. This and subsequent Biennial Conferences, which are designed to provide for systematists of all kinds, included themed symposia that resulted in further publications. The Association also publishes volumes that are not specifically linked to meetings and encourages new publications in a broad range of systematics topics.

Anyone wishing to learn more about the Systematics Association and its publications should refer to our website at www.systass.org.

Other Systematics Association publications are listed after the index for this volume.

The Systematics Association Special Volume Series 72

Reconstructing the Tree of Life

Taxonomy and Systematics of Species Rich Taxa

Edited by

Trevor R. Hodkinson

John A. N. Parnell

Department of Botany
School of Natural Sciences
Trinity College Dublin
Dublin, Ireland

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an informa business

Outside to inside of image: water ermine moth, UK (*Spilosoma urticae*); barley, UK (*Hordeum distichon*); fossilised sea urchins, Tunisia (*Mecaster* spp.); seeds, unknown origin (Bignoniaceae); purple sea snails, worldwide (*Janthina janthina*); fossilised shark teeth, USA (*Isurus* sp.); and sea urchin, Greece (*Arbacia lixula*)

artwork by: Diccon Alexander (diccona@hotmail.com)

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-10: 0-8493-9579-8 (Hardcover)
International Standard Book Number-13: 978-0-8493-9579-6 (Hardcover)

Library of Congress Cataloging-in-Publication Data

Reconstructing the tree of life : taxonomy and systematics of species rich taxa / editors, Trevor R. Hodkinson and John A.N. Parnell.
p. cm. -- (The Systematics Association special volume series)
Includes bibliographical references and index.
ISBN 0-8493-9579-8 (alk. paper)
1. Biology--Classification. I. Hodkinson, Trevor R. II. Parnell, John A. N.

QH83.R43 2006
578.01'2--dc22 2006048341

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Preface

The twenty chapters of this book are based on the theme of the plenary session of the Fourth Biennial Conference of the Systematics Association, held at Trinity College Dublin (TCD), Ireland, in August 2003, namely the systematics of species rich taxa. During the five-day conference, there were stimulating presentations, posters and discussions, covering a broad sample of the 'tree of life'; these also influenced the shape and content of this volume. Papers were contributed by a number of conference delegates and by others subsequently invited to broaden the book's scope or address particular theoretical issues.

Consideration of the book's theme and content began at a conference planning meeting at TCD in early 2003 with the local conference organiser, Steve Waldren of TCD, and Gordon Curry, the honorary treasurer of the Systematics Association. These were refined further in discussions with Alan Warren, the Systematics Association special volumes series editor, and Chris Humphries, the president of the Systematics Association. We are grateful to all of them for their input and encouragement, particularly our colleague, Steve. Two anonymous book proposal reviewers also provided valuable content guidance. We are particularly grateful for the manuscript preparation input of Sandra Velthuis of Whitebarn Consulting, who has worked long and hard to proofread chapters and standardise their format, and to the production team, especially Gail Renard, Pat Roberson and John Sulzycki, at CRC Press, who have been highly supportive and professional. We also thank Diccon Alexander for the superb cover artwork. Finally we thank all 51 contributing authors to the book, many of whom also peer reviewed other chapters. We encourage all readers to support the activities of the Systematics Association (www.systass.org).

Trevor R. Hodkinson
John A.N. Parnell
Department of Botany
School of Natural Sciences
Trinity College Dublin
Ireland

The Editors

Dr Trevor Hodkinson is Senior Lecturer in the Department of Botany, School of Natural Sciences, Trinity College Dublin (TCD), Ireland. He is head of the Molecular Laboratory and specialises in the research fields of molecular systematics, genetic resources and taxonomy (http://www.tcd.ie/Botany/Staff/THodkinson.html).

Professor John Parnell is also from the Department of Botany at TCD. He is curator of the herbarium and his research interests are mainly in the fields of taxonomy and systematics (http://www.tcd.ie/Botany/Staff/JParnell.html).

Contributors

T.G. Barraclough
Division of Biology and NERC Centre for Population Biology
Imperial College London, UK

E. Biffin
Division of Botany and Zoology
Australian National University
Canberra, Australia

O.R.P. Bininda-Emonds
Institut für Spezielle Zoologie und Evolutionsbiologie mit Phyletischem Museum
Friedrich-Schiller-Universität Jena
Jena, Germany

K.E. Black
School of Forest Resources
Penn State University
University Park, Pennsylvania, USA

Y. Bouchenak-Khelladi
Department of Botany
School of Natural Sciences
Trinity College Dublin
Dublin, Ireland

J. Brodie
Botany Department
The Natural History Museum
London, UK

G. Cassis
Research and Collections Branch
Australian Museum
Sydney, Australia

M.W. Chase
Jodrell Laboratory
Royal Botanic Gardens, Kew
Richmond, UK

J.J. Clarkson
Jodrell Laboratory
Royal Botanic Gardens, Kew
Richmond, UK

J.A. Cotton
Zoology Department
The Natural History Museum
London, UK

L.A. Craven
Australian National Herbarium
Centre for Plant Biodiversity Research
Canberra, Australia

C.J. Creevey
European Molecular Biology Laboratory
EMBL Heidelberg
Heidelberg, Germany

T.J. Davies
Department of Biology
University of Virginia
Charlottesville, Virginia, USA

R.P.J. de Kok
Herbarium
Royal Botanic Gardens, Kew
Richmond, UK

D.A. Fitzpatrick
Conway Institute
University College Dublin
Dublin, Ireland

G. Fusco
Department of Biology
University of Padova
Padova, Italy

M. Geerts
Burg. Heynenstraat 11
Swalmen, The Netherlands

K.W. Hilu
Department of Biological Sciences
Virginia Polytechnic Institute and State University
Blacksburg, Virginia, USA

T. R. Hodkinson
Department of Botany
School of Natural Sciences
Trinity College Dublin
Dublin, Ireland

K.D. Hyde
Centre for Research in Fungal Diversity
Department of Ecology and Biodiversity
The University of Hong Kong
Hong Kong, China

S.W.L. Jacobs
National Herbarium
Royal Botanic Gardens
Sydney, Australia

M.S. Kinney
Department of Botany
School of Natural Sciences
Trinity College Dublin
Dublin, Ireland

A.F. Konings
Cichlid Press
El Paso, Texas, USA

J.O. McInerney
Department of Biology
National University of Ireland Maynooth
Maynooth, Ireland

K.R. McKaye
Appalachian Laboratory
University of Maryland System
Frostburg, Maryland, USA

A. Minelli
Department of Biology
University of Padova
Padova, Italy

E. Negrisolo
Department of Public Health, Comparative Pathology and Veterinary Hygiene
University of Padova
Legnaro, Italy

M.J. O'Connell
Department of Biochemistry
University College Cork
Cork, Ireland

J.A.N. Parnell
Department of Botany
School of Natural Sciences
Trinity College Dublin
Dublin, Ireland

G. Petersen
Botanical Garden and Museum
The Natural History Museum of Denmark
Copenhagen, Denmark

D.E. Pisani
Department of Biology
National University of Ireland Maynooth
Maynooth, Ireland

G. Reid
Botany Department
The Natural History Museum
London, UK

N. Rønsted
Jodrell Laboratory
Royal Botanic Gardens, Kew
Richmond, UK

N. Salamin
Department of Ecology and Evolution
University of Lausanne
Lausanne, Switzerland

V. Savolainen
Jodrell Laboratory
Royal Botanic Gardens, Kew
Richmond, UK

F.R. Schram
Department of Biology
University of Washington
Seattle, Washington, USA

R.T. Schuh
American Museum of Natural History
Division of Invertebrate Zoology
New York, New York, USA

O. Seberg
Botanical Garden and Museum
The Natural History Museum of Denmark
Copenhagen, Denmark

B.D. Shenoy
Centre for Research in Fungal Diversity
Department of Ecology and Biodiversity
The University of Hong Kong
Hong Kong, China

A. Stamatakis
Swiss Federal Institute of Technology
School of Computer and Communication Sciences
Lausanne, Switzerland

J.R. Stauffer, Jr.
School of Forest Resources
Penn State University
University Park, Pennsylvania, USA

M. Steel
Biomathematics Research Centre
University of Canterbury
Christchurch, New Zealand

A.M.C. Tang
Centre for Research in Fungal Diversity
Department of Ecology and Biodiversity
The University of Hong Kong
Hong Kong, China

K. Turk
Jodrell Laboratory
Royal Botanic Gardens, Kew
Richmond, UK

T.M.A. Utteridge
Herbarium
Royal Botanic Gardens, Kew
Richmond, UK

M.A. Wall
Department of Entomology
San Diego Natural History Museum
San Diego, California, USA

W.C. Wheeler
Division of Invertebrate Zoology
American Museum of Natural History
New York, New York, USA

M. Wilkinson
Zoology Department
The Natural History Museum
London, UK

D.M. Williams
Botany Department
The Natural History Museum
London, UK

E. Yektaei-Karin
Jodrell Laboratory
Royal Botanic Gardens, Kew
Richmond, UK

G.C. Zuccarello
School of Biological Sciences
Victoria University of Wellington
Wellington, New Zealand

Contents

SECTION A Introduction and General Context

SECTION B Reconstructing and Using the Tree of Life

Section A

Introduction and General Context

1 Introduction to the Systematics of Species Rich Groups

T. R. Hodkinson and J. A. N. Parnell
Department of Botany, School of Natural Sciences, Trinity College Dublin

CONTENTS

ABSTRACT

To completely document the world's diversity of species we need to undertake some simple but mountainous tasks; above all we need to tackle its species rich groups. We need to collect them, name and classify them, and position them on the tree of life. We need to do this systematically across all groups of organisms, and because of the biodiversity crisis we need to do it quickly. A qualitative approach to defining a species rich taxon — such as a species rich genus, family, order, class or phylum — appears more broadly applicable than a quantitative definition, but combining such categories of definition also appears useful. We define a species rich group as: 'a group with a relatively high number of species in comparison to other groups of the same, and comparable, taxonomic rank'. This chapter introduces, with examples, the concept of species rich groups and discusses how these groups are central to efforts to document the world's diversity of species and to help address the biodiversity crisis. Naming and describing species rich groups is the first step in placing them on the phylogenetic tree of life. Phylogenetic trees are becoming bigger (supersized) and methods are being developed to deal with the computational complexity of such trees. This paper also outlines the wider context of the book and papers presented herein. With species rich taxa, evolution has set taxonomists and systematists a difficult, but not unattainable, challenge that must be addressed as a matter of urgency.

1.1 INTRODUCTION

It may be a surprise to many readers that biologists cannot answer two seemingly simple yet fundamental questions: 'how many species are there in the world?' and 'how do the world's species relate to one another in an evolutionary context?'. The first question is a basic challenge for taxonomists who list, describe and classify the world's organisms. The second is a challenge for systematists/phylogeneticists who try to place organisms in an evolutionary framework by inferring a tree of life such as that shown in Figure 1.1. Activities of both groups of workers are critically impeded by species rich taxa, as they are often poorly sampled and described, yet make up a high proportion of total global species richness.

There is a huge variance in the published estimates of the total number of species on Earth. It could lie anywhere in the region of 4 million to 100 million[1–3]. We cannot even accurately count the number of species that have so far been described because of synonymy (the same species unwittingly recorded under different names by different researchers, that is, duplication). For example, 1.7 million species have been described but levels of synonymy could be in the range of 20–50%[4–6] (but see Cassis et al., *Chapter 13*, for a higher value). Even for a particular species rich group, estimates can vary enormously. For example, in the insects with approximately 1 million described species, estimates of the total number of species have varied from 1.8 million by Hodkinson and Casson[7] to 80 million by Stork[8]. An intermediate 10 million, proposed by Ødegaard et al.[9], may well be more appropriate, but such estimates are often based on crude methods (Cassis et al., *Chapter 13*). Furthermore, for many species rich groups, only a low proportion of the total estimated number of species has been described. For example, approximately 100,000 fungi have been described but 1.5 million species may exist (Tang et al., *Chapter 15*), only 15,000–20,000 diatom species (heterokont algae) have been described but up to 200,000 may exist (Williams and Reid, *Chapter 19*) and approximately 5,800 red algae (Rhodophyta) have been described but 20,000 may exist (Brodie and Zuccarello, *Chapter 20*).

Why do estimates of the number of species in the world vary by an order of magnitude or more, and why is there such uncertainty? Some of the reasons are covered in the chapters of this book, particularly *Chapter 2* (Schram) and *Chapter 3* (Seberg and Petersen), but one problem stands out above all others, namely that of the species rich groups. It is probably fair to say that taxonomists have collected representatives of most of the major lineages (groups) of life and that the discovery of new major branches is a rare event meriting high publicity; for example that surrounding a new species, *Symbion pandora,* discovered feeding on the mouth of the Norway lobster and assigned to a new phylum, Cycliophora[10]. However, there is now a need to fill in the gaps to find and characterise, in an evolutionary framework, all the other representatives belonging to those groups and particularly, in the context of this book, its species rich taxa.

Species diversity is not evenly distributed across the range of life forms that have existed on Earth. If species were distributed evenly between and within major groups of organisms, and if the taxonomic units were strictly comparable, we could simply and accurately count the number of species in one section of the tree (Figure 1.2a) and multiply up by the number of comparable sections so that the whole tree is represented. However, this pattern is not seen in nature, and we find striking examples of imbalance. Some evolutionary lineages have succeeded while others have perished. For example the hexapods, a group including the insects, are a species rich group compared to their closest relatives the myriapods, crustaceans, cheliceriformes and tardigrades (Figure 1.2b) and all other eukaryotic life (Cassis et al., *Chapter 13*). Furthermore, there may be as many as 200,000 diatoms (heterokont algae), but their sister group has recently been recognised as a group of tiny flagellates, Bolidophyceae, which has no more than three to five currently recognised species[11,12] (see also Williams and Reid, *Chapter 19*). Therefore, speciation and extinction are not random processes; some groups of organism have speciated to a staggering degree, while others have not. The factors leading to such imbalance are discussed throughout this book but especially in *Chapter 10* (Davies and Barraclough), *Chapter 11* (Hilu) and *Chapter 17* (Hodkinson et al.).

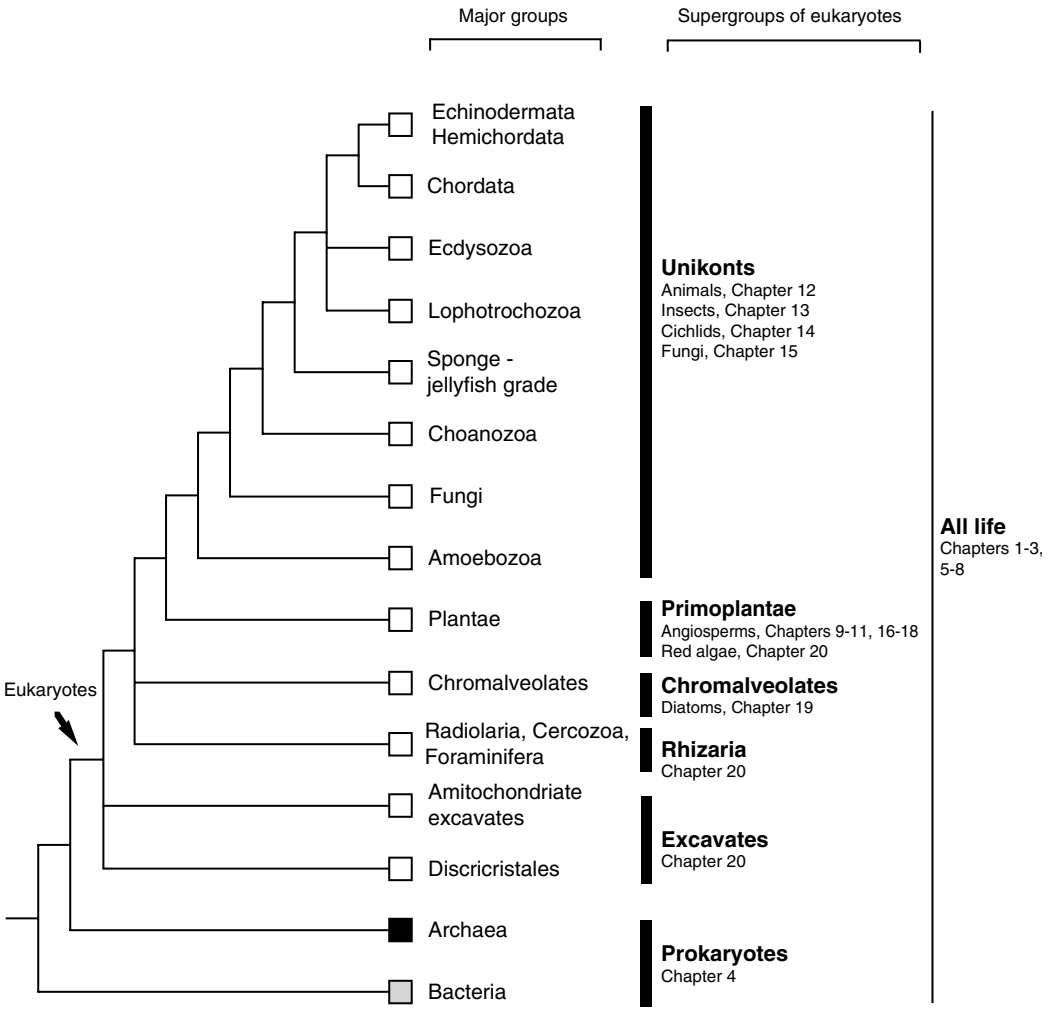


FIGURE 1.1 Tree of life. Chapters within the book that relate to specific species rich taxa are indicated. Open squares represent eukaryotes, the black square represents archaea and the hatched square represents bacteria. Representatives of the major groups include (1) *Bacteria:* hydrogenobacteria, blue-green bacteria, green-sulphur bacteria, spirochaetes; (2) *Archaea:* korarchaeotes, crenarchaeotes, euryarchaeotes; (3) *Discricristales:* euglenids, trypanosomes, acrasid slime moulds; (4) *Amitochondriate excavates:* parabasalids, diplomonads; (5) *Radiolaria:* radiolarians; (6) *Cercozoa:* cercomonads; (7) *Foraminifera:* foraminiferans; (8) *Chromalveolates:* diatoms, brown algae, oomycetes (water moulds), ciliates, dinoflagellates; (9) *Plantae:* angiosperms (flowering plants), gymnosperms, ferns, liverworts, mosses, green algae; (10) *Amoebozoa:* slime moulds, lobose amoebae (mycetozoans); (11) *Fungi:* microsporidians, zygomycetes, basidiomycetes, ascomycetes; (12) *Choanozoa:* choanoflagellates, ichthyosporeans; (13) *Sponge — jellyfish grade:* siliceous 'sponges', calcareous 'sponges', corals, jellyfish, aceolomorphs; (14) *Lophotrochozoa:* gastropods (snails), bivalves (clams), platyhelminths, rotifers, brachiopods; (15) *Ecdysozoa:* nematodes, insects, centipedes, crabs, barnacles, spiders, velvet worms; (16) *Chordata:* humans, birds, lizards, fish, lancelets, tunicates; (17) *Echinodermata:* sea urchins, sea cucumbers; and (18) *Hemichordata:* acorn worms. (Major groups and representatives adapted from Pennisi[2] and supergroups of eukaryotes from Baldauf[27].)

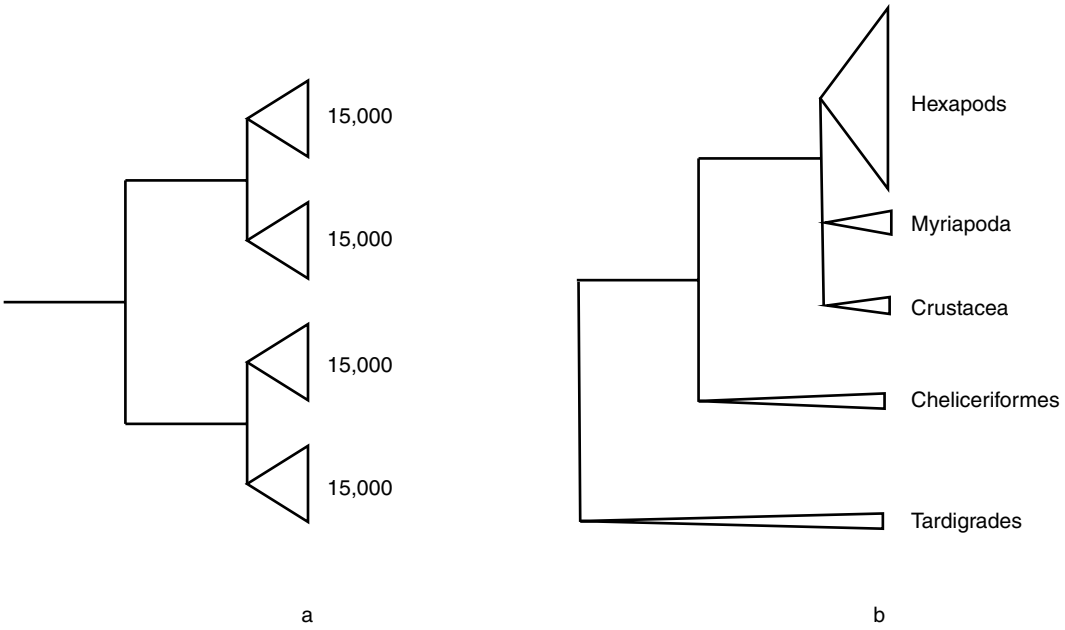


FIGURE 1.2 Species richness of phylogenetic groups is not evenly distributed. (a) If speciation and extinction had proceeded in a stochastic manner we would not expect to see significant levels of variation from the model shown (in a fully resolved and bifurcating tree). Triangles are drawn in proportion to species richness in that clade (15,000 species in all clades of Figure 1.2a). (b) An example of imbalance in species diversification within the animal group comprising the insects. Insects belong to the hexapods and account for three quarters of all described animal diversity. The hexapod clade is much larger in terms of species number than any of its sister groups of same taxonomic rank (Mriapoda, Crustacea and Cheliceriformes). (Figure 1.2b adapted from Cassis et al., *Chapter 13*.)

This book concerns the taxonomy and systematics of species rich groups; it is about how to collect, document, describe and classify them. It is also about the inextricably linked phylogenetic studies that try to position species rich taxa on the tree of life and represent their diversity. This introduction defines species rich groups, highlights examples of major species rich groups, introduces the concept of the tree of life and discusses the problems and prospects of dealing with species rich groups. It unashamedly focuses on species rich groups. Species poor groups are obviously important components of world species diversity, but they lie outside the aims and scope of this volume.

1.2 WHAT IS A SPECIES RICH GROUP?

1.2.1 Quantitative and Objective Definitions

Surprisingly, there is little literature on what the essential properties of a species rich group are, nor much discussion of how such groups might be defined. Rather it seems assumed that a species rich group will always be easily and universally recognisable as such and therefore needs no formal definition. We disagree and believe that it is important to attempt to define what constitutes a species rich group. Such a definition could be quantitative or qualitative, or both. In a quantitative or objective approach we might try to define a species rich group in everyday numerical terms. We could, for example, simply give a numerical threshold, which the size of a group must exceed, before it can be classified as species rich or 'big'. Frodin[13] takes this approach for plant genera and defines a 'big genus' as one containing 500 or more species. The same argument could be applied at different taxonomic ranks; that is, we could identify suitable thresholds that could be

considered as big. For example, a large family could be defined as containing at least 5,000 species or a large order as containing at least 20,000 species. This approach may work within some groups such as the angiosperms or insects and for comparisons between them. For example, the grass family (Hilu, *Chapter 11* and Hodkinson et al., *Chapter 17*) and the insect bug family Miridae (Cassis et al., *Chapter 13*) both contain approximately 10,000 species and both can be considered, under this definition, to be species rich families. These families can also be considered big in that they usually present a mountainous challenge to systematists specialising in the group.

Therefore, a quantitative approach can sometimes work, but it soon runs into difficulties if used in a wider context. For example, the threshold value given above could not be used to sensibly describe the largest families of mammal because no mammal families or genera would be considered big under such a definition; there are only an estimated 5,500 mammal species in approximately 1,000 genera which themselves tend to be small. The largest mammal order, Rodentia, contains 2,000–3,000 species, but the largest mammal family, Muridae (including mice, rats and gerbils), has approximately 600 species[14–17]. Likewise this threshold figure could not be used for the fish suborder, Labroidei, containing the cichlids (Stauffer et al., *Chapter 14*), a group with approximately 1,800 species. Clearly this is unsatisfactory, as the cichlids, in most biologists' minds, are species rich (850 species of cichlid have been found in the African Great Lake Malawi alone).

A further complication in trying to numerically define a species rich group is that there are no quantitative ways of defining a particular taxonomic rank. Taxonomic ranks are clearly defined in a relative hierarchical sense (a genus is a collection of species; a family a collection of genera, and so on) but not in any absolute numerical sense. Without such common yardsticks, taxonomists can recognise species and classify them in different ways, and because of this, a taxonomic group in one rank does not necessarily represent the same degree of distinction (evolutionary divergence) as that in another taxonomic group of the same rank. For this reason it is often not possible to make meaningful comparisons from one taxonomic group to another even if they are from the same rank.

The size of taxonomic groups can also be quantified using a phylogenetic approach and sister clade comparisons. A clade may be large in comparison to its sister clade(s). For example, Hexapoda in Figure 1.2b are much more species rich than Myriapoda and Crustacea. This approach allows us to get a relative measure for comparative purposes but is not widely applicable beyond the sister clades in question. For example, both Myriapoda and Crustacea can be considered large in comparison to many other animal groups of the same taxonomic rank. This quantitative method is also open to the same problems of transferability between taxonomic groups as is the basic quantitative definition of a species rich group discussed above. Thresholds must be chosen in order to say how big a group has to be to be regarded as species rich in comparison to its sister groups.

1.2.2 Qualitative and Subjective Definitions

If we recognise that 'big' for one group is 'small' for another, then we may prefer a qualitative (that is, relative or subjective) definition of a species rich group. A species rich group could therefore be defined as: 'a group with a relatively high number of species in comparison to other groups of the same, and comparable, taxonomic rank'. The caveat 'comparable' has been added to the definition to avoid the problem introduced by the wide taxonomic comparisons discussed above and the lack of common yardsticks in taxonomy.

Clearly using a qualitative approach such as that suggested above immediately leads to the well known 'hollow curve' of Willis[18,19] discussed by many other authors (including Hilu, *Chapter 11,* and to a lesser degree Parnell et al., *Chapter 16*). Therefore, to some extent we are here entering the realm of Dial and Marzluff[20] who argued that an index of dominance (the ratio of N_{Max}/N_{Tot}, where N_{Max} is the number of subtaxa in the largest taxon and N_{Tot} is the total number of subtaxa) could be used to characterise the size distributions of taxa. Clearly, as this index is not dependent on the absolute value of N_{Tot}, high values of the index are comparable across different taxonomic groups and so could be used to define a taxon rich group. But what value of the index should be chosen? We further discuss

TABLE 1.1
Top Five Species Rich Orders of Insects

Orders	Species	% of All Insect Species
Coleoptera (beetles)	350,000	35.0
Lepidoptera (butterflies and moths)	150,000	15.0
Hymenoptera (bees and wasps)	125,000	12.5
Diptera (flies)	120,000	12.0
Hemiptera (true bugs, cicadas, leafhoppers, aphids)	90,000	9.0
Total	835,000	83.5

Note: The five largest orders, representing 6.4% percent of all insect orders, contain approximately 83.5% of all insect species.

Source: Cassis et al., *Chapter 13*, and references therein.

the hollow curve in Section 1.2.3 below, where we attempt to combine quantitative and qualitative definitions, and in Section 1.3.4, where patterns and processes are tackled. The following examples serve to illustrate the qualitative definition of species rich groups, namely the species rich insects and the species rich angiosperms and various subgroups within each.

Within the species rich hexapods (Cassis et al., *Chapter 13;* Figure 1.2b), the insects dominate, and so far approximately 1 million species have been described and divided into 31 orders. Insects also make up approximately three quarters of all animal species that have been described. The insects are, therefore, clearly species rich hexapods and species rich animals. Within the insects, the vast majority of species are found in one of five orders (Coleoptera, Diptera, Hymenoptera, Lepidoptera and Hemiptera). These represent 835,000 of the species and over 80% of all insect species diversity (Table 1.1). They can without difficulty be called species rich orders. The top five families account for 21% of the species (Table 1.2) despite representing less than 1% of all insect families, and 20 insect families account for almost 45% of the insects; these can all legitimately be termed species rich families. A number of species rich genera can also be identified, such as *Agrilus* (Coleoptera) with over 8,000 species, *Camponotus* (Hymenoptera) with over 1,500 species, and *Megaselia* (Diptera) also with over 1,500 species (Wall, personal communication).

Such a pattern of uneven species distribution holds true across all major groups of life. For example, within the angiosperms (more than 250,000 species in 13,185 genera[21]), five families

TABLE 1.2
Top Five Species Rich Families of Insects

Families	Species	% of All Insect Species
Curculionidae (weevils and snout beetles)	50,000	5.4
Staphylinidae (rove beetles)	47,000	5.1
Cerambycidae (long horned beetles)	35,000	3.8
Chrysomelidae (leaf beetles, flea beetles, root worms)	35,000	3.8
Carabidae (ground beetles)	30,000	3.2
Total	197,000	21.3

Note: The five largest families (all beetles), representing less than 1% of all insect families, contain approximately 21% of all insect species.

Source: Cassis et al., *Chapter 13*, and references therein.

TABLE 1.3
Top Five Species Rich Families of Angiosperms

Families	Species	% of All Species	Genera	% of All Genera
Asteraceae (daisies)	22,750	9.1	1,528	11.6
Orchidaceae (orchids)	18,500	7.4	788	6.0
Fabaceae (beans)	18,000	7.2	624	4.7
Rubiaceae (coffees)	10,200	4.1	630	4.8
Poaceae (grasses)	9,500	3.8	668	5.1
Total	78,950	31.6	4,238	32.1

Note: The largest five families, representing just 1% of all angiosperm families, contain 31.6% of all angiosperm species.

Source: Data from Mabberley[21].

(beans, coffees, daisies, grasses, orchids) account for 31.6% of the species and 32.1% of the genera (Table 1.3), so these can be legitimately defined as species rich families. The top 10 angiosperm genera all have more than 1,000 species and account for 7% of all angiosperm species, the largest 15 for 9.3% (Table 1.4) and the largest 50 for 19.8% (data not shown) despite representing only

TABLE 1.4
Top 15 Species Rich Angiosperm Genera

Rank	Genus (Family)	Number of Species
1	*Astragalus* (Fabaceae)	3,270
2	*Bulbophyllum* (Orchidaceae)	2,032
3	*Psychotria* (Rubiaceae)	1,951
4	*Euphorbia* (Euphorbiaceae)	1,836
5	*Carex* (Cyperaceae)	1,795
6	*Begonia* (Begoniaceae)	1,484
7	*Dendrobium* (Orchidaceae)	1,371
8	*Acacia* (Fabaceae)	1,353
9	*Solanum* (Solanaceae)	1,250
10	*Senecio* (Asteraceae)	1,250
11	*Croton* (Euphorbiaceae)	1,223
12	*Pleurothallis* (Orchidaceae)	1,120
13	*Eugenia* (Myrtaceae)	1,113
14	*Piper* (Piperaceae)	1,055
15	*Ardisia* (Myrsinaceae)	1,046
	Total	23,149

Note: All top 10 angiosperm genera have at least 1,000 species, and together they contain 7% of the angiosperm species despite only representing 0.075% of the genera. The top 15 largest genera contain 23,149 species (9.3% of all angiosperms) despite representing 0.1% of all angiosperm genera. *Syzygium* (Mrytaceae) ranks 16th with 1,041 species (but see Parnell et al., *Chapter 16*), and *Ficus* ranks 31st with 750 species and is the topic of *Chapter 9* (Rønsted et al.).

Source: Figures from Frodin[13] and percentages calculated from total angiosperm species and genus numbers in Mabberley[21].

0.4% of all angiosperm genera (calculated from values given in Frodin[13] and Mabberley[21]). These can all be defined as species rich genera. The taxonomy and systematics of two species rich angiosperm genera are explored in more detail within this book; *Syzygium* with between 1,000 and 1,500 species (Parnell et al., *Chapter 16*) and *Ficus* with 750 species (Rønsted et al., *Chapter 9*). Whilst we believe that the pattern of uneven distribution does allow for the construction of a qualitative definition of a species rich group it is somewhat unsatisfactory in that it is largely subjective. How are the defining percentages to be set?

1.2.3 Combining Objective and Subjective Definitions

It is clear that both quantitative and qualitative definitions are, to some extent, problematic. However, it is possible to combine these categories and to define a species rich group using a combination of qualitative and quantitative criteria.

As shown in *Chapter 11* (Hilu) and *Chapter 16* (Parnell et al.), the distribution of subtaxa within taxa follows a hollow curve distribution. Of the taxa so distributed, this volume is concerned with those that are large relative to the rest. Cronk[22] pointed out the asymmetry of the size distributions of taxa and indicated that the variance of the lognormal distribution is the best general descriptor of the hollow curve. Scotland and Sanderson[23] compared a number of hypothetical distributions with curves generated from real data. They concluded that none of their tested hypothetical distributions matched their real hollow curves very well. However, as only four real hollow curves were tested, the transferability of their conclusions remains unclear. In this book, we seek to define species rich groups and differentiate them from groups that are not species rich. In other words, we need to define where to stop as we slide down the hollow curve towards its tail; we require a stopping rule (or some other way of deciding how to partition the curve). Jackson's discussion of stopping rules applicable for ecological ordination[24] is, we believe, of relevance, although his favoured solutions cannot be applied, partly because of the findings of Scotland and Sanderson[23]. However, it appears to us that an extension of the concept of the scree plot, despite its disadvantages[24], does offer an opening first approximation to a definition satisfactorily combining subjective and objective criteria.

The idea underlying use of the scree plot in the context of Jackson[24] is that there is a break point in a curve, where its slope flattens out, and that values in the flat part of the curve may be disregarded. In our case, we are interested not in this particular point on the curve but in the concept of a break point. In particular, is there a break point in the tail of the curve; that is, is the tail continuous or fragmented towards its tip? Examination of the tails of a number of published hollow curves or of the data used to construct them, generally does show a break towards the end of the tail. For example, there are break points visible in tails of all the curves published by Scotland and Sanderson[23]. We understand that there may well be cases where the break point is not obvious nor possibly even singular. However, in general for the curves and data we have seen, there does appear to be an obvious gap. For example, in the case of Myrtaceae (Parnell et al., *Chapter 16*) our method yields three species rich genera in the family — *Syzygium, Eucalyptus* and *Eugenia.* Such a procedure seems to yield a relatively small number of truly exceptionally taxa that are exceptionally subtaxa rich; other workers may wish to extend the concept of species rich groups further into the flat part of the curve, perhaps defining species rich groups in terms of the uppermost quartile or some other non-arbitrary concept.

1.2.4 Large Taxonomic Groups

The basis of the concept for a qualitative (or combined qualitative and quantitative definition) of a species rich group can be extended to other large taxonomic assemblages. For this reason we can distinguish between 'large taxa' and 'subtaxa rich taxa'. Not all large groups are subtaxa rich, but most will be. By recognising other large taxonomic groups we accommodate the taxonomic

hierarchy. For example, a family can be considered a large group because it contains a large number of genera, with no reference to the number of species. So we can use terms such as 'genus rich family', 'family rich order' or 'order rich class'.

1.3 RECONSTRUCTING AND USING THE TREE OF LIFE

1.3.1 The Tree of Life

Naming all the world's species is just a first step in understanding them. We need to know how each of these organisms relates to one another and how they are positioned on the tree of life. The tree of life model is one of the most enduring and powerful tools at the disposal of an evolutionary biologist. It is a hypothesis or statement of inferred relationships between organisms displayed in a graphical form approaching a branching, tree-like structure. Such trees, also known as phylogenetic trees, have evolved from the early attempts of Charles Darwin[25] and Ernst Haeckel[26] to more sophisticated and presumably accurate trees such as Figure 1.1. This tree has been divided up into seven major subgroups (Unikonts, Primoplantae, Chromalveolates, Rhizaria, Excavates, Archaea and Bacteria, following Baldauf[27]) and 15 minor subgroups (following Pennisi[2]). For more detailed trees, with more subgroups, see Pennisi[2], Baldauf[27], Cracraft and Donoghue[28], and Palmer et al.[29].

Despite the power of phylogenetic trees, there has been much debate about whether a tree of life exists and whether life can be accurately represented using a phylogenetic tree model or a combination of trees[30–34] (see also McInerney et al., *Chapter 4*). The answer is yes and no. Life is unlikely to be fully represented by an all-encompassing and unambiguous tree of life model. Endosymbiosis, genome fusion, horizontal gene transfer, hybridisation, polyploidy and reticulation are all substantive issues that sometimes make simple trees unrealistic approximations of phylogeny[33,35] (see also Rønsted et al., *Chapter 9*). However, the evolution of life is likely to have taken, at least within eukaryotes, a tree-like pattern (Figure 1.1 and Figure 1.3a). Within most eukaryotes there is little reason to suggest that such processes occur commonly enough to prevent the recovery of a tree of life except possibly near some reticulating tips[29]. Steel (*Chapter 7*) adds weight to this argument by showing, with a simple mathematical result, that an underlying tree of life can always be defined (and exists) even in the presence of complications such as reticulation. He shows how the notion of a tree of life can be rigorously defended but recognises that such a tree defined in the presence of complications such as reticulation will miss much of the detail and richness of evolutionary history and will be largely unresolved in places. He goes on to discuss methods for better representing and studying reticulate evolution. However, within the prokaryotes horizontal gene transfer and genome fusions have been more common, and this begs the question of whether an underlying tree structure exists and is recoverable[29,32] (see also McInerney et al., *Chapter 4*).

There are numerous other models that can be used to describe the evolutionary history of organisms, and some of these are particularly apt for prokaryotes where horizontal gene transfer and endosymbiosis have played a more significant role in evolution. They may also be more appropriate for closely related species where frequent hybridisation is known to occur. Therefore, it is clear that a basic three-domain tree of all life is an oversimplification[30], a network of some sort with a 'universal ancestor', or network of gene trees, may better explain the pattern[31,32] (see also Steel, *Chapter 5*). Zimmer[36] has coined this concept the 'mangrove of life' (Figure 1.3b). A more recent concept to emerge is the 'ring of life' (Rivera and Lake[34]). This concept has the potential to represent prokaryotic evolution and the origin of eukaryotes. Rivera and Lake's ring of life (Figure 1.3c) is based on the analysis of hundreds of genes and a method called 'conditioned reconstruction' that uses shared genes as a measure of genome similarity and allows horizontal gene transfer to be used in assessing genome based phylogeny. It resolves the dual nature of eukaryotic genomes that sit simultaneously on an eubacterial lineage (bacteria) and an archaebacterial lineage (archaea). This is what seals the ring (Martin and Embley[33]). Their model supports

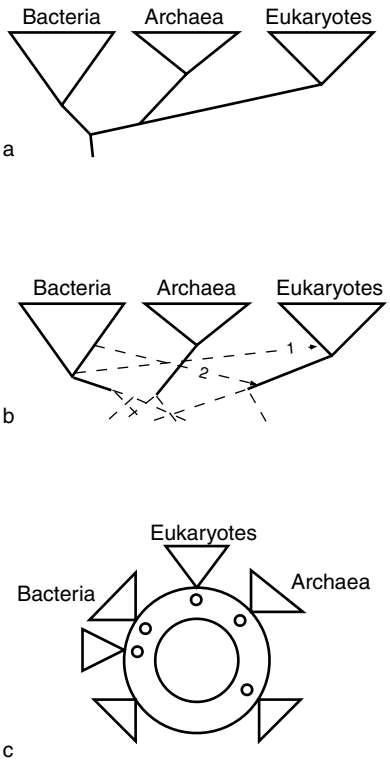


FIGURE 1.3 Tree of life models. (a) A standard phylogenetic tree showing the three domains of life; within the triangles a standard tree like branching pattern is seen. (b) A network tree incorporating reticulation; reticulations are seen by endosymbiotic events (fusion of genomes) and by exchange of genes in gene trees. For example one event involved bacteria giving rise to chloroplasts (1) and another event involved bacteria giving rise to mitochondria (2). (c) A ring of life, a model used to depict evolutionary pattern especially useful for the prokaryotes and origin of the bigenomic eukaryotes. Small circles within the ring represent defining ancestors of the major groups. (Figure 1.3a adapted from Woese[30], 1.3b from Zimmer[36]; 1.3c from Rivera and Lake[34].)

the idea that a union has occurred between achaebacterial and eubacterial genomes, likely to be an endosymbiotic association between two prokaryotes. The evolution of prokaryotes and the notion of a prokaryotic tree are discussed further in McInerney et al. (*Chapter 4*).

The debate about the shape of the tree, network or ring of life is essential and stimulating. However, all models are by definition imperfect, but many are good enough to work from, and a simple tree is as good a place to start as anywhere else (as it is the simplest of the models). Even though a network or other model may better explain these patterns, they may not have the same analytical power or simplicity of a tree (or combination of trees).

1.3.2 Big Tree Reconstruction for Species Rich Groups: Are Large Phylogenetic Trees Accurate?

Most phylogenetic studies have included relatively few species, and only a few studies have included the large numbers of taxa required for detailed understanding of species rich groups or other large tree of life problems[37,38]. The next decade will see the rise of supersized phylogenetic trees (Hodkinson et al., *Chapter 17*) because DNA sequencing has become a standard laboratory technique and costs have dropped. Advances in DNA sequencing techniques are also envisaged. Phylogenetic analyses will therefore include more characters and more species.

One major concern is whether methods of phylogenetic reconstruction can accommodate large datasets. The first step in the production of phylogenetic trees often involves applying a method of phylogenetic reconstruction such as maximum likelihood, parsimony analysis or Bayesian inference. The second step in maximum likelihood and parsimony (but not Bayesian inference) involves the assessment of internal support, via resampling methods such as bootstrapping and the jacknife[39,40] so that the investigator can discriminate between groups with clear phylogenetic signals and those needing more investigation or more data to resolve[44]. The production of large phylogenetic trees and assessing internal support via resampling methods are mathematical and computational challenges because they involve searches of tree space (the total set of possible trees for the relevant set of taxa), and the number of possible trees grows more than exponentially with the number of taxa on the tree. This means that, as the number of taxa increases, the job of accurately finding the optimal trees under some objective function becomes relatively much more difficult due to the increase in tree space. We must therefore ask whether existing methods can, or will ever be able to, accurately reconstruct the phylogeny of species rich groups with several thousands and possibly hundreds of thousands of taxa. These are the topics explored by Wilkinson and Cotton (*Chapter 5*), Bininda-Emonds and Stamatakis (*Chapter 6*) and Steel (*Chapter 7*) and to a lesser extent by Wheeler (*Chapter 8*) and Hodkinson et al. (*Chapter 17*). Despite the scale of the problem there is cause for optimism. Increasing the number of characters in a dataset[42–44] and the number of taxa sampled[45–47] (see also Hodkinson et al., *Chapter 17*) generally results in more reliable phylogenetic inferences, if not limited significantly by computational issues. At some point the computational complexity of the problem must, however, outweigh the benefits of adding taxa (Bininda-Emonds and Stamatakis, *Chapter 6*).

Empirical and theoretical studies show that existing methods perform relatively well with large datasets[47–49]. For example, Salamin et al.[44] have shown, using Monte Carlo simulations, good accuracy of parsimony and neighbour joining methods to retrieve model trees with taxon numbers up to 13,000 (the number of angiosperm genera and close to the number of species in a large angiosperm family such as the grasses) if sequences of sufficient length (number of nucleotides) were used (see Hodkinson et al., *Chapter 17*). Testing the reliability of phylogenetic inference using, for example, resampling methods is also a major challenge with large DNA matrices[41]. However, existing methods and shortcuts perform relatively well[41], and we expect that advances in tree search methods will facilitate this process.

Better and more powerful phylogenetic methods are being developed and tested for analysing large computationally demanding phylogenetic datasets. These methods can be categorised into supermatrix and supertree methods[50–52] (see also Bininda-Emonds and Stamatakis, *Chapter 6;* Steel, *Chapter 7*). Supermatrix and supertree approaches are not mutually exclusive, as supertrees are essential in many formal divide-and-conquer analysis methods of single datasets (supermatrices). These divide-and-conquer strategies seek to break down the problem into smaller subproblems (a process known as decomposition) that are computationally easier to solve (Wilkinson and Cotton, *Chapter 5*). The results from these subproblems are then combined to provide an answer for the initial global problem. Large analyses may incorporate divide-and-conquer search strategies such as quartet puzzling and disk covering. These methods are likely to become increasingly important for analyses of large data sets as well as for searches of smaller data sets using more complex and computationally demanding optimality criteria.

Wilkinson and Cotton (*Chapter 5*) discuss advances in supertree methodology as part of a divide-and-conquer strategy. They explore the issue of effective taxon overlap and how it may be achieved via suitable decomposition, and they present a new fast supertree method. Bininda-Emonds and Stamatakis (*Chapter 6*) further discuss theoretical issues surrounding the reconstruction of large phylogenetic trees. They investigate the potential to reconstruct phylogenies for species rich groups and ever-larger portions of the tree of life using a range of methods; they explore the scalability of phylogenetic accuracy with respect to species number. Their results show that taxon number itself, especially with the implementation of disk covering methods, may not be the

constraining factor in these analyses but that the strategy used to sample taxa may have a larger impact on both accuracy and analysis time.

1.3.3 Characters and Homology

Accurate phylogenetic analysis is critically based on the input of high quality phylogenetically informative characters (that is, 'good' characters), and these can be of many types but are predominantly molecular, morphological or anatomical. Obviously, different types of data are useful for the study of different evolutionary processes and at different levels of evolutionary divergence/taxonomic rank. For example, nucleotide substitutions within single-locus nuclear genes are proving highly valuable for studies of closely related species. Likewise, combinations of different genes including nuclear, plastid and mitochondrial genomes are utilised for studies of hybridisation, introgression and polyploidy in such closely related species (discussed in depth by Rønsted et al., *Chapter 9*). Morphological characters are essential for many analyses including those of extinct fossil species[29] and are vital in investigations of evolution and development (evo-devo). Minelli et al. (*Chapter 12*) outline the importance of morphology in evo-devo studies and show how it can help with phylogenetic reconstruction in general.

The use of DNA sequence data in phylogenetic analysis requires assumptions to be made about the homology of characters (positional homology of nucleotides within aligned sequences). This is often an overlooked problem and is particularly important for analyses of large datasets. Wheeler (*Chapter 8*) explores this critical issue of homology assessment and describes the various solutions to the problem. Sequence availability is also an issue and is discussed in Hodkinson et al. (*Chapter 17*).

1.3.4 Patterns and Processes of Diversity and Understanding the Hollow Curve

Large phylogenetic trees can be used for the study of pattern and processes in evolution but also a whole list of other biological questions. Dobzhansky's statement that 'nothing in biology makes sense except in the light of evolution'[52] has almost become a cliché but remains highly relevant and pertinent.

One of the most commonly used applications of large trees is for classification and taxonomy. However, they also have wider application to a host of biological and evolutionary questions[53,54]. Large trees have convenience from a statistical perspective (Steel, *Chapter 7*) and there are many theoretical reasons for using large trees[46,55–57]. For example, they are required for accurate inferences of macro-evolutionary processes because in such studies it is desirable to sample most of the diversity within a study group to reduce the risk of incorrect phylogenetic tree reconstruction and to allow meaningful comparisons to be made or hypotheses to be tested[40,53].

Large phylogenetic trees of species rich taxa are useful tools for detecting diversification rate variation, extinction and exploring the processes that may have led to the diversity of the group. We may, for example, wish to know why some groups have become species rich and others have either failed to diversify or have perished. The distribution of species richness within a phylogenetic tree, even between closely related groups of organisms, can vary enormously.

As discussed above, the hollow curve[18,19] has been used to describe patterns of diversification where few taxonomic groups are species rich while the majority are species poor. There may be, for example, an inverse relationship of large to small genera (that is, lots of small genera and few large ones). Within the angiosperms the frequency distribution of genera containing increasing numbers of species (number of species in a genus plotted against the number of genera) approximates to the logarithmic hollow curve, although the first term is always larger than expected. Because of this, classifications are generally strongly polarised, having some 80% of the genera smaller than average but some 80% of the species concentrated in genera larger than average[58,59]. Age of a genus, species richness of genera and geographical area that the genus occupies tend to

be correlated, although there are opposing views as to how that correlation maps out. Cronk[22] considers large genera to be recent blooms of evolution, whereas Willis interpreted big genera as being old (for further discussion see Hilu, *Chapter 11*). Modern phylogenetic reconstruction allows these alternative hypotheses to be tested. Widespread genera are often larger than continental genera[60,61]. Clayton and Renvoize[61] suggest that there may be a dichotomy in evolutionary strategies between large genera speciating in a wide variety of niches and small genera in labile environments subject to continuing processes of disruption and replacement. These are hypotheses that require detailed analysis and testing. The properties of the hollow curve and processes leading to it are discussed in detail by Hilu (*Chapter 11*) and Parnell et al. (*Chapter 16*).

A number of tests using the temporal and/or topological properties of phylogenetic trees exist to determine if diversification variation is statistically significant[62–65]. In the species rich angiosperms, for example, diversification can vary by over several orders of magnitude between clades (Davies and Barraclough, *Chapter 10*). Furthermore, within any particular angiosperm family, such as the grasses, diversification rates have also been shown to vary (Hodkinson et al., *Chapter 17*). Factors including key biological traits, coevolution, geography and environmental variables may have contributed to the variation that exists in net diversification between clades[62,65]. Davies and Barraclough (*Chapter 10*) review studies to explore diversification in flowering plants using large scale phylogenetic trees. They also discuss further statistical tests to explore these processes. Rønsted et al. (*Chapter 9*) also discuss the coevolution and cospeciation of *Ficus* with hymenopteran wasps belonging to the species rich insect family Agonidae.

1.4 TAXONOMY OF SPECIES RICH GROUPS

If we are going to document and understand the diversity of species in the world, and that of species rich groups in particular, we need to make sure that some basic tasks, including the collecting, naming, describing and classifying of those organisms, are undertaken. We need to complete these tasks systematically across all groups of organisms, and because of the currently high rate of ablation of biodiversity (the biodiversity crisis), we need to complete these tasks soon. We are facing a potentially massive episode of extinction, so it is essential that such studies are carried out as quickly as possible so that conservation policies and strategies are based on the best possible information.

1.4.1 Collecting

Collecting trips need to avoid unnecessary duplication and ensure that the maximum species diversity is sampled. They also need to be shown to be good value for money. Collecting is one of the main rate determining steps in documenting the world's species and further characterising them. The topics of how we should focus and prioritise our collecting efforts to maximise new species discovery are covered in *Chapter 18* (Utteridge and de Kok) and to a lesser degree in *Chapter 2* (Schram) and *Chapter 3* (Seberg and Petersen). Collecting is a slow and expensive process. For example, over 100 grasses were collected in a recent two-week period in New South Wales and Queensland, Australia, by the first author of this chapter and Surrey Jacobs, a highly cooperative and experienced grass taxonomist, from the Royal Botanic Gardens, Sydney. One of the grass species, *Alexfloydia repens,* is only known from one location in the world. The second, *Homopholis belsonii,* is very rare and endangered (Jacobs, personal communication). Both species took close to a day to track down and collect, entailing considerable financial expense, not to mention leech attacks, tick infestations and mosquito bites (bloody biodiversity!). Beyond such anecdotal statements, others have tried to quantify the pace of collecting in an attempt to estimate the scale of the task. Parnell's quantification of the costs of collecting[66] showed that about 85% of the costs of collecting a specimen for a number of expeditions were salary associated, with 63% being direct salary costs. Surprisingly, he showed that expenses such as travel, local living and

postage for a collecting expedition, which is the part external agencies are most likely to be asked to fund (and without which the expedition simply cannot occur), constituted only about 12–17% of the total costs. Seberg and Petersen (*Chapter 3*) and Cassis et al. (*Chapter 13*) have tried to quantify the effort required to sample species rich groups by doing some simple calculations based on the number of people days it will take to collect all remaining species of a species rich group.

Such estimates allow us to see the scale and potential cost of the problem, but we should also remember this is only part of the process. It covers the resources required for collection, but not the additional resources needed for describing and classifying the organisms (that could amount to the same or more again). In reality these figures are also likely to be underestimates because geographical areas will need to be resampled many times, at different times of the year, with different methods (with specialist and generalist collectors; see Utteridge and de Kok, *Chapter 18*) before we can be sure that we are close to collecting all species in an area.

1.4.2 Naming, Describing and Classifying

The process of naming, describing and classifying organisms is sometimes known as alpha taxonomy (Williams and Reid, *Chapter 19*); it is time consuming and requires highly qualified staff. For some taxa, the shortage of specialists is an issue, leading to huge delays in identification. Therefore, ensuring some degree of evenness of taxonomic coverage is an important issue. Taxonomy needs to be done across the board, not just for well known organisms (we have provided examples in the latter section of this book for a range of taxonomic groups). The focus tends to be on well known groups and may be excessive. Working on the relatively small genus *Cyclamen* (c. 30 species) Compton et al.[67] indicate that the 'differing infrageneric classifications produced in *Cyclamen* result from varying taxon sampling, differing interpretation of morphological data, changes in the sources and analysis of data, and inconsistent application of names'. They conclude that 'extensive subdivision of small genera in the absence of adequate data that could provide evidence for consistent patterns of relationship is premature and leads to a proliferation of names'. Clearly, large or species rich genera offer far more potential for inappropriate subdivision, a topic briefly discussed in Parnell et al. (*Chapter 16*). Concentration on relatively well known groups may occur at the expense of the less well known ones, a real problem if those less well known groups are also big, a topic discussed in Schram (*Chapter 2*).

The taxonomic coverage of papers in this book spans the tree of life and can be seen in Figure 1.1. *Chapter 2* (Schram) and *Chapter 3* (Seberg and Petersen) introduce general issues, and more specific discussions are given for insects (Cassis et al., *Chapter 13*), fish (Stauffer et al., *Chapter 14*), fungi (Tang et al., *Chapter 15*), angiosperms (Rønsted et al., *Chapter 9;* Parnell et al., *Chapter 16;* Hodkinson et al., *Chapter 17*), diatoms (Williams and Reid, *Chapter 19*) and algae (Brodie and Zuccarello, *Chapter 20*). Many of the chapters discuss the advances made in electronic resources that make 'taxonomic information readily available at the click of a mouse' (Bisby et al.[68]). Such systems will involve 'terascale taxonomy', having to handle enormous volumes of information including data, literature and images, on behaviour, classification, ecology, genome, geography, morphology, nomenclature, ontogeny, phylogeny and physiology (Wheeler et al.[69]). Considering an estimated world species number of 10 or more million, this will ultimately result in trillions of observations associated with specimens in natural history collections[69]. Digital databasing has started[68] and is making good progress. It will certainly facilitate taxonomic work and make information globally available by linking institutions such as museums, herbaria, universities and their taxon specialists. For specialist species rich groups there are several existing high quality database systems that can be used as models (Schram, *Chapter 2*). There is therefore, as Schram explains, no need to reinvent the wheel, although a review of such systems could be useful. Experiences with some model groups such as the plant bugs (Cassis et al., *Chapter 13*) and grasses (Hodkinson et al., *Chapter 17*) should be evaluated and recommendations made on how best develop other

systems. We must also remember that the digital interface is only a tool and cannot replace well trained taxonomists or physical resources such as herbaria and museums. These resources will only work with international cooperation. Such coordinated action at an international level is also needed to reach consensus over taxonomic nomenclature and accepted names.

The DNA revolution has offered huge potential to taxonomy and systematics, but as with the digital revolution, we should take care. Obviously we should be prepared to embrace the methods where they can offer real help. For example, a recent development that may help with the taxonomy and systematics of species rich groups is DNA barcoding and DNA taxonomy. The slow pace of species description and taxonomy has led some to call for a modern DNA based taxonomy[70–72]. In this method, DNA sequences are used to identify the organism. Sequences are generated and compared to sequences found in a database that have known identity and are linked to real, accurately identified specimens in institutions such as herbaria and museums. The appeal of this fully automated approach is that anybody should be able to identify an organism without specialist knowledge of the group. It also offers the potential to develop futuristic tools that can instantaneously identify an organism by sampling its DNA and making a comparison to a database of sequences. This would have particular advantages in species rich groups where taxon identification is often a problem and synonymy a big issue.

However, there are a number of issues with this technology, especially if interpreted in the strict sense, including concerns about sequence quality, insufficient sampling within and amongst species, pseudogenes, herbarium specimen quality and availability, type specimen use and common occurrence of hybridisation and introgression and associated DNA exchange (capture) between closely related species. Seberg and Petersen (*Chapter 3*) discuss the pitfalls of DNA technology and highlight the danger of using it inappropriately as a shortcut in taxonomy. DNA barcoding is seen by many as a better alternative in that it uses DNA sequences to aid identification but is not all prevailing when it comes to identification. DNA can certainly facilitate and improve taxonomy. DNA sequences have the added bonus that they have high potential for phylogenetics, classification and for providing a phylogenetic framework for developing a meaningful monographic study (Hodkinson et al., *Chapter 17*), although caveats may apply[73]. Phylogenetics, molecular systematics and taxonomy are therefore inextricably linked.

1.5 CONCLUSIONS: BLAME EVOLUTION AND POLITICIANS

This book is concerned with species rich groups. By concentrating on such groups we do not mean to suggest that species poor groups should be ignored. Far from it, but they are outside the scope of this book. We divide this book into three sections:

- Introduction and general context
- Reconstructing and using the tree of life
- Taxonomy and systematics of species rich groups (case studies)

To document and characterise the world's species rich groups is one of the largest challenges of biology and needs financial and political support. The reason this challenge has not been adequately addressed is partly because evolution has set us an enormous task and partly because politicians have not prioritised the problem sufficiently highly; we should therefore blame both evolution and politicians. However, the task is achievable. Schram, in the next chapter, outlines his vision of how this could be achieved. Readers may not agree with all his points but will hopefully find some common ground on most of them. It will require the meshing together of phylogenetics and taxonomy, considerable advances in informatics, improved and increased collecting, training of taxonomists and significant financial support. We hope that this book goes some way to help achieve that aim.

ACKNOWLEDGEMENTS

We thank Gerry Cassis, Nicolas Salamin and Michael Wall for comments on this manuscript.

REFERENCES

1. Blackmore, S., Biodiversity update: progress in taxonomy, *Science,* 298, 365, 2002.
2. Pennisi, E., Modernizing the tree of life, *Science,* 300, 1692, 2003.
3. Wheeler, Q.D., Taxonomic triage and the poverty of phylogeny, *Phil. Trans. R. Soc. Lond. B,* 359, 571, 2004.
4. Gaston, K.J. and May, R.M., The taxonomy of taxonomists, *Nature,* 356, 281, 1992.
5. Solow, A.R., Mound, L.A., and Gaston, K.J., Estimating the rate of synonymy, *Syst. Biol.,* 44, 93, 1995.
6. May, R.M., The dimensions of life on earth, in *Nature and Human Society: The Quest for a Sustainable World,* National Academy of Sciences Press, Washington DC, 2000.
7. Hodkinson, I.D. and Casson, D., A lesser predilection for bugs — Hemiptera (Insecta) diversity in tropical rain-forests, *Biol. J. Linn. Soc.,* 43, 101, 1991.
8. Stork, N.E., Insect diversity: facts, fiction and speculation, *Biol. J. Linn. Soc.,* 35, 321, 1988.
9. Ødegaard, F., Diserud, O.H., and Ostbye, K., The importance of plant relatedness for host utilization among phytophagous insects, *Ecol. Lett.,* 8, 612, 2005.
10. Funch, P. and Kristensen, R.M., Cycliophora is a new phylum with affinities to Entoprocta and Ectoprocta, *Nature,* 378, 711, 1995.
11. Guillou, L. et al., *Bolidomonas:* a new genus with two species belonging to a new algal class, the Bolidophyceae (Heterokonta), *J. Phycol.,* 35, 368, 1999.
12. Kühn, S., Medin, M., and Eller, G., Phylogenetic position of the parasitoid nanoflagellate *Pirsonia* inferred from nuclear-encoded small subunit ribosomal DNA and a description of *Pseudopirsonia* n. gen. and *Pseudopirsonia mucosa* (Drebes) comb. nov., *Protist,* 155, 143, 2004.
13. Frodin, D.G., History and concepts of big plant genera, *Taxon,* 53, 753, 2004.
14. Vaughan, T.A., Ryan, J.M., and Capzaplewski, N.J., *Mammalogy,* 4th ed., Saunders College Publishing, 2000.
15. Michaux, J., Reyes, A., and Catzeflis, F., Evolutionary history of the most speciose mammals: molecular phylogeny of muroid rodents, *Molec. Biol. Evol.,* 17, 280, 2001.
16. O'Leary, M.A. et al., Building the mammalian sector of the tree of life: combining different data and a discussion of divergence times for placental mammals, in *Assembling the Tree of Life,* Cracraft, J. and Donoghue, M.J., Eds., Oxford University Press, Oxford, 2004, 490.
17. Wilson, D.E. and Reeder, D.M., Eds., *Mammal Species of the World,* 3rd ed., Johns Hopkins University Press, 2005.
18. Willis, J.C., *Age and Area,* Cambridge University Press, Cambridge, 1922.
19. Willis, J.C., The birth and spread of plants, *Boissera,* 8, 1949.
20. Dial, K.P. and Marzluff, J.M., Nonrandom diversification within taxonomic assemblages, *Syst. Zool.*, 38, 26, 1989.
21. Mabberley, D.J., *The Plant Book,* 2nd ed., Cambridge University Press, Cambridge, 1997.
22. Cronk, Q., Measurement of biological and historical influences on plant classifications, *Taxon,* 38, 357, 1989.
23. Scotland, R.W. and Sanderson, M.J., The significance of few versus many in the tree of life, *Science,* 303, 643, 2004.
24. Jackson, D.A., Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches, *Ecology,* 74, 2204, 1993.
25. Darwin, C., *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life,* John Murray, London, 1859.
26. Haeckel, E., *Generale Morphologie der Organismen,* Verlag von Georg Reimer, Berlin, 1866.
27. Baldauf, S.L., The deep roots of eukaryotes, *Science,* 300, 1703, 2003.
28. Cracraft, J. and Donoghue, M.J., *Assembling the Tree of Life,* Oxford University Press, Oxford, 2004.
29. Palmer, J.D., Soltis, D.E., and Chase, M.W., The plant tree of life: an overview and some points of view, *Amer. J. Bot.,* 91, 1437, 2004.

30. Woese, C.R., Kandler, O., and Wheelis, M.C., Towards a natural system of organisms: proposal for the domains Archaea, Bacteria and Eucarya, *Proc. Natl. Acad. Sci. USA,* 87, 4576, 1990.
31. Woese, C.R., The universal ancestor, *Proc. Natl. Acad. Sci. USA,* 95, 6854, 1998.
32. Doolittle, W.F., Phylogenetic classification and the universal tree, *Science,* 284, 2124, 1999.
33. Martin, W. and Embley, T.M., Early evolution comes full circle, *Nature,* 431, 134, 2004.
34. Rivera, M.C. and Lake, J.A., The ring of life provides evidence for a genome fusion origin of eukaryotes, *Nature,* 431, 152, 2004.
35. Linder, C.R. and Rieseberg, L.H., Reconstructing patterns of reticulate evolution in plants, *Amer. J. Bot.,* 91, 1700, 2004.
36. Zimmer, C., *Evolution: The Triumph of an Idea,* William Heinemann, London, 2002, 101.
37. Savolainen, V. and Chase M.W., A decade of progress in plant molecular phylogenetics, *Trends Genet.,* 19, 717, 2003.
38. Sanderson, M.J. and Driskell, A.C., The challenge of constructing large phylogenetic trees, *Trends Plant Sci.,* 8, 374, 2003.
39. Efron, B., Bootstrap methods: another look at the jackknife. *Ann.*, *Stat.,* 7, 1, 1979.
40. Felsenstein, J., Phylogenies and the comparative method, *Am. Nat.,* 125, 1, 1985.
41. Salamin N., et al., Assessing internal support with large phylogenetic DNA matrices, *Molec. Phylogenet. Evol.,* 27, 528, 2003.
42. Erdos, P.L. et al., A few logs suffice to build (almost) all trees: part II, *Theor. Comp. Sci.,* 221, 77, 1999.
43. Bininda-Emonds, O.R.P., et al., Scaling of accuracy in extremely large phylogenetic trees, in *Pacific Symposium on Biocomputing 6,* Altman, R.B., et al., Eds., World Scientific Publishing Company, River Edge, New Jersey, 2001, 547.
44. Salamin, N., Hodkinson T.R., and Savolainen, V., Towards building the tree of life: a simulation study for all angiosperm genera, *Syst. Biol.,* 54, 183, 2005.
45. Hillis, D.M., Inferring complex phylogenies, *Nature,* 383, 130, 1996.
46. Hillis, D.M., Taxonomic sampling, phylogenetic accuracy, and investigator bias, *Syst. Biol.,* 47, 3, 1998.
47. Källersjö, M. et al., Simultaneous parsimony jackknife analysis of 2538 *rbcL* DNA sequences reveals support for major clades of green plants, land plants, seed plants and flowering plants, *Pl. Syst. Evol.,* 213, 259, 1998.
48. Soltis, P.S., Soltis, D.E., and Chase, M.W., Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology, *Nature,* 402, 402, 1999.
49. Savolainen, V. et al., Phylogeny reconstruction and functional constraints in organellar genomes: plastid versus animal mitochondrion, *Syst., Biol.,* 51, 638, 2002.
50. Salamin, N., Hodkinson T.R., and Savolainen, V., Building supertrees: an empirical assessment using the grass family (Poaceae), *Syst. Biol.,* 51, 136, 2002.
51. Wilkinson, M. et al., The shape of supertrees to come: tree shape related properties of fourteen supertree methods, *Syst. Biol.*, 54, 419, 2005.
52. Dobzhansky, T., Nothing in biology makes sense except in the light of evolution, *Am. Biol. Teach.*, 35, 125, 1973.
53. Purvis, A., Using interspecies phylogenies to test macroevolutionary hypotheses, in *New Uses for New Phylogenies,* Harvey, P.H. et al., Eds., Oxford University Press, Oxford, 1996, 153.
54. Harvey, P.H. et al., Eds., *New Uses for New Phylogenies,* Oxford University Press, Oxford, 1996.
55. Rannala, B. et al., Taxon sampling and the accuracy of large phylogenies, *Syst. Biol.,* 47, 702, 1998.
56. Källersjö, M., Albert, V.A., and Farris, J.S., Homoplasy increases phylogenetic structure, *Cladistics,* 15, 91, 1999.
57. Hillis, D.M. et al., Is sparse taxon sampling a problem for phylogenetic inference? *Syst. Biol.,* 52, 124, 2003.
58. Clayton, W.D., Some aspects of the genus concept, *Kew Bull.,* 27, 281, 1972.
59. Clayton, W.D., The logarithmic distribution of angiosperm families, *Kew Bull.,* 29, 271, 1974.
60. Clayton, W.D., Chorology of the genera of Gramineae, *Kew Bull.*, 30, 111, 1975.
61. Clayton, W.D. and Renvoize, S.A., *Genera Graminum: Grass Genera of the World,* Her Majesty's Stationery Office, London, 1986.
62. Barraclough, T.G. and Nee, S., Phylogenetics and speciation, *Trends Ecol. Evol.,* 16, 391, 2001.
63. Chan, K.M.A. and Moore B.R., Whole-tree methods for detecting differential diversification rates, *Syst. Biol.,* 51, 855, 2002.

64. Chan, K.M.A. and Moore B.R., SYMMETREE: whole-tree analysis of differential diversification rates*, Bioinformatics,* 21, 1709, 2004.
65. Moore, B.R., Chan, K.M.A., and Donoghue, M.J., Detecting diversification rate variation in supertrees, in *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life,* Bininda-Emonds, O.R.P., Ed., Kluwer Academic Publishers, Dordrecht, 2004, 487.
66. Parnell, J.A.N., The monetary value of herbarium collections*,* in *Biological Collections and Biodiversity,* Rushton, B.S., Hackney, P., and Tyrie, C.R., Eds., Linnean Society of London Special Publication 3, England, 2001, 271.
67. Compton, J.A., Clennett, J.C.B., and Culham, A., Nomenclature in the dock. Overclassification leads to instability: a case study in the horticulturally important genus *Cyclamen, Bot. J. Linn. Soc.,* 146, 339, 2004.
68. Bisby, F.A. et al., Taxonomy, at the click of a mouse, *Nature,* 418, 367, 2002.
69. Wheeler, Q.D., Lipscomb, D., and Platnick, N., Terascale taxonomy: cyber-infrastructure and the Linnaean legacy, in *Proc. of the Fourth Biennial Conference of the Systematics Association,* Trinity College Dublin, Ireland, 2003.
70. Tautz, D. et al., DNA points the way ahead in taxonomy, *Nature,* 418, 479, 2002.
71. Tautz D. et al., A plea for DNA taxonomy, *Trends Ecol. Syst.,* 18, 70, 2003.
72. Lipscomb, D., Platnick N., and Wheeler, Q., The intellectual content of taxonomy: a comment on DNA taxonomy, *Trends Ecol. Syst.,* 18, 65, 2003.
73. Stace, C.A., Plant taxonomy and biosystematics: does DNA provide all the answers? *Taxon,* 54, 999, 2005.

2 Taxonomy/Systematics in the Twenty-First Century

F. R. Schram
Department of Biology, University of Washington, Seattle, USA
Formerly of Zoological Museum, University of Amsterdam, The Netherlands

CONTENTS

And out of the ground the Lord God formed every beast of the field, and every fowl of the air; and brought them unto Adam to see what he would call them: and whatsoever Adam called every living creature, that was the name thereof. And Adam gave names to all cattle, and to the fowl of the air, and to every beast of the field … (Genesis 2:19–20)

ABSTRACT

Taxonomy/systematics has had a history extending back to the 1880s, with Cassandras issuing dire warnings about the future of the science, but little hard data exist to document these warnings. Some institutions have done well, while others have endured severe cutbacks or even disappeared. Meanwhile, the need for effective biodiversity knowledge is increasing exponentially. The numbers of species in many groups is truly staggering, and the use of information technology to manage terascale volumes of data in the science of taxonomy is inarguably essential. The tools to effectively move on this need to be developed, and online models for specific groups of organisms including