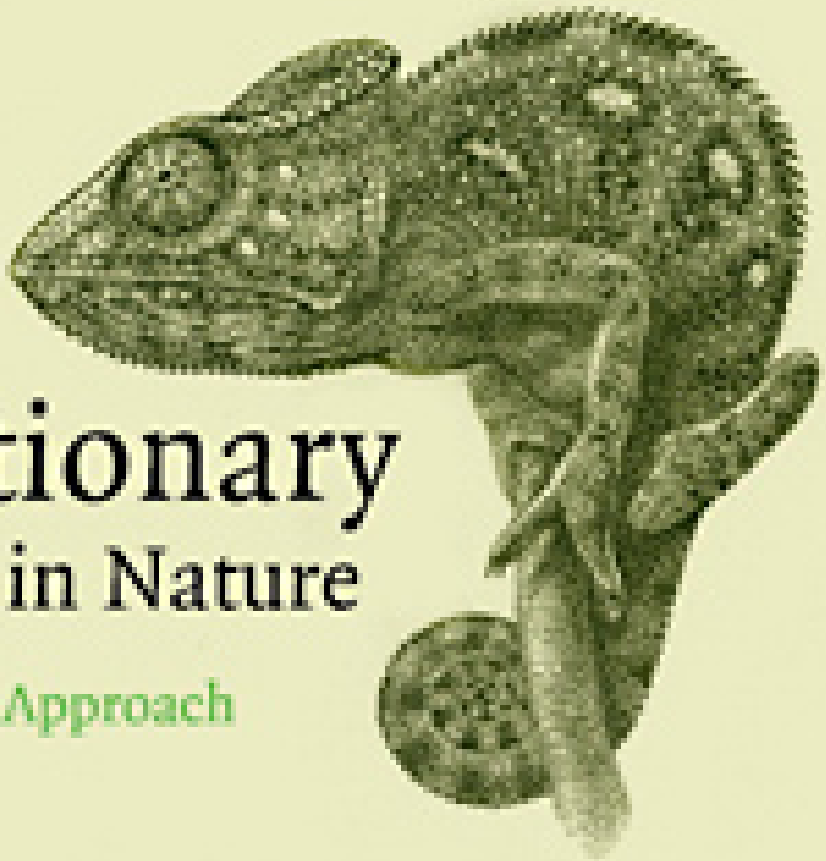

Evolutionary Pathways in Nature

A Phylogenetic Approach

JOHN C. AVISE

This page intentionally left blank

Evolutionary Pathways in Nature

A Phylogenetic Approach

Reconstructing phylogenetic trees from DNA sequences has become a popular exercise in many branches of biology, and here the award-winning geneticist John Avise explains why. Molecular phylogenies provide a genealogical backdrop for interpreting the evolutionary histories of many other types of biological traits (anatomical, behavioral, ecological, physiological, biochemical, and even geographical). Guiding readers on a natural history tour along dozens of evolutionary pathways, the author describes how creatures ranging from microbes to elephants came to possess their current phenotypes. If you want to know how the toucan got its bill and the kangaroo its hop, then this is the book for you. This book also provides a definitive answer to the proverbial question: 'which came first, the chicken or the egg?' This scientifically educational yet entertaining treatment of ecology, genetics and evolution is intended for college students, professional biologists, and anyone interested in natural history and biodiversity.

John C. Avise is a distinguished professor of Ecology and Evolutionary Biology at the University of California, Irvine.

Evolutionary Pathways in Nature

A Phylogenetic Approach

JOHN C. AVISE
Department of Ecology and Evolutionary Biology
University of California

Illustrations by Trudy Nicholson

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 2RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521857536

First published in print format 2006

ISBN-13 978-0-511-21789-0 eBook (Adobe Reader)
ISBN-10 0-511-21789-7 eBook (Adobe Reader)

ISBN-13 978-0-521-85753-6 hardback
ISBN-10 0-521-85753-8 hardback

ISBN-13 978-0-521-67417-1 paperback
ISBN-10 0-521-67417-4 paperback

CONTENTS

Preface *page* ix
Acknowledgments xi

1 Introduction 1
The meaning of phylogeny 1
Phylogenetic metaphors 3
Molecular appraisals of phylogeny 8
Comparative phylogenetics 11
Phylogenetic character mapping 15

2 Anatomical structures and morphologies 19
Whence the toucan's bill? 19
The beak of the fish 22
Snails' shell shapes 25
More on snails' shell shapes 28
Winged walkingsticks 30
Hermits and kings 34
True and false gharials 36
Loss of limbs on the reptile tree 39
Fishy origins of tetrapods 42
Panda ponderings 44
Fossil DNA and extinct eagles 47
The Yeti's abominable phylogeny 48

3 Body colorations 52
Light and dark mice 52
Sexual dichromatism 54
Dabbling into duck plumages 58
Specific avian color motifs 62
The poisonous Pitohui 65

Warning colorations in poison frogs 68
Müllerian mimicry butterflies 71
Caterpillar colors and cryptic species 73

4 Sexual features and reproductive lifestyles 77
The chicken or the egg? 77
The avian nest 80
Egg dumping and foster parentage 84
Egg laying and live bearing 87
Piscine placentas 90
Male pregnancy 93
Living and reproducing by the sword 95
Brood care in Jamaican land crabs 98
Social parasitism of butterflies on ants 101
Of monkeyflowers and hummingbirds 104
Parthenogenetic lizards, geckos, and snakes 107
Delayed implantation 110

5 More behaviors and ecologies 114
The kangaroo's bipedal hop 114
Powered flight in winged mammals 118
Magnetotaxis in bacteria 121
Cetacean origins 122
Feeding and echolocation in whales 124
The phylogeny of thrush migration 129
Pufferfish inflation 131
Eusociality in shrimp 134
Evolutionary reversals of salamander lifecycles 137
Dichotomous life histories of marine larvae 140
Adaptive radiations in island lizards 143
Spiders' web-building behaviors 145
Lichen lifestyles 150

6 Cellular, physiological, and genetic traits 153
Foregut fermentation 153
Snake venoms 156
Antifreeze proteins in anti-tropical fish 159
Warm-bloodedness in fishes 162
Electrical currents 166

The Xs and Ys of sex determination 168
The eyes have it 170
Two types of body 174
The phylogenomics of DNA repair 178
Roving nucleic acids 180
Host-to-parasite gene transfer 184
Tracking the AIDS virus 187

7 Geographical distributions 190
Afrotheria theory 190
Aussie songbirds 193
Madagascar's chameleons 196
The evolutionary cradle of humanity 200
Coral conservation 204
Sri Lanka, a cryptic biodiversity hotspot 207
Overseas plant dispersal 210
Phylogenetic bearings on Polar Bears 212
Looking over overlooked elephants 215
Bergmann's rule 218

Epilog 221
Appendix: a primer on phylogenetic character mapping 223
History of cladistic concepts and terminology 223
Maximum parsimony 228
Maximum likelihood 231
Independent contrasts between pairs of quantitative traits 233
Glossary 239
References and further reading 253
Index 279

PREFACE

Many biologists now incorporate molecular phylogenetic analyses into their explorations of nature. Using sophisticated laboratory techniques, they uncover "DNA markers" or "genetic tags" that uniquely identify each creature. Furthermore, details in the submicroscopic structures of these natural labels offer tantalizing clues to how living organisms were genealogically linked through bygone ancestors. Thus, lengthy DNA sequences housed in the cells of all organisms carry not only the necessary molecular genetic instructions for life, but also extensive records of phylogeny, i.e. of evolutionary ancestry and descent.

During the replication and transmission of DNA from one generation to the next, mutations continually arise. Many of these spread through populations (via natural selection, or sometimes by chance genetic drift), thereby cumulatively altering particular molecular passages in each species' hereditary script. In recent years, scientists have learned how to read and interpret the genealogical content of these evolutionary diaries – these "genomic autobiographies" – of nature. Results are summarized as phylogenetic diagrams that depict how particular forms of extant life are connected to one another via various historical branches in the Tree of Life.

Phylogenetic analysis has become a wildly popular exercise in many areas of biology, but phylogenies estimated from DNA sequences are seldom the ultimate objects of scientific interest. The primary value of each molecular phylogeny lies instead in its utility as historical backdrop for deciphering the evolutionary histories of other kinds of biological traits such as morphologies, physiologies, behaviors, lifestyles, or geographical distributions. By mapping such organismal features onto species' phylogenies estimated from molecular data, biologists can address fascinating questions of the following sort. Did the bipedal hop arise once or multiple times in kangaroo evolution? From what type of ancestor did toucan birds evolve their banana-like bills? How often during evolution have reptiles lost their limbs? Are the antifreeze proteins in Arctic and Antarctic fishes functionally similar by virtue of shared ancestry or convergent evolution? By what evolutionary routes have some fishes evolved powerful electrical discharges? Did Jamaican land crabs derive their peculiar forms of offspring care from a common ancestor? Did

walkingstick insects evolve from flyingsticks or vice versa, and how often? How have certain bacteria acquired their magnetic compasses? On how many occasions have distinct algal and fungal lineages joined forces in lichen symbioses? Where on the planet have phylogenetic appraisals uncovered cryptic species and conservation-relevant hotspots of global biodiversity? Can the ancient breakup of the supercontinent Gondwanaland account for the modern distributions of particular lineages of birds and mammals in the Southern Hemisphere? Where and when did the viruses responsible for the AIDS epidemic enter the human species? And, which came first: the chicken or the egg?

By highlighting studies that have provided scientific answers to these and many additional questions, I intend to illustrate the power (and also some limitations) of comparative phylogenetic perspectives in biological research. Several available textbooks describe, in depth, how molecular data are gathered in the laboratory and analyzed at the computer. My approach here will not be to recount the many operational details of molecular phylogenetics (although introductory background is provided). Rather, my intent is to serve as a naturalist guide on a biological expedition into the remarkable world of nature, as viewed through the evolutionary prism of molecular phylogeny. In each of 67 essays arranged into six topical chapters, I describe how a DNA-estimated phylogeny provided historical framework for interpreting a puzzling ecological feature or evolutionary process in organisms with unusual anatomies or lifestyles, or in creatures with special significance to one or another biological field such as ethology, natural history, biogeography, conservation, biochemistry, physiology, epidemiology, or medicine.

Through this case-history approach, I hope to provide a fun yet educational introduction – for amateur naturalists and students to professional biologists – to how comparative phylogenetic analyses have helped to solve some of nature's most intriguing mysteries. Another goal is to encourage a deeper appreciation of the many intellectual and aesthetic treasures of the biological world. As more and more people become educated about nature's ways, perhaps societies will learn to cherish life's variety and strive harder to preserve what remains. Tragically, through human actions, populations and species today are being driven to extinction at rates seldom experienced in the planet's long history. To terminate any lineage now is to lose forever a genetic wisdom that was honed along an epic evolutionary journey lasting nearly four billion years. Paradoxically, life is both fragile and tenacious. Extinction continually threatens, and once realized can never be undone. However, having withstood and adapted to countless environmental challenges over the geological eons, each extant lineage is also a hardy and proven survivor, surely deserving of our deepest respect and admiration.

ACKNOWLEDGMENTS

Doug Futuyma, Blair Hedges, David Hillis, Kirk Jensen, Judith Mank, Axel Meyer, David Reznick, DeEtte Walker, John Ware, and several anonymous reviewers provided helpful comments on various portions of the text. I am especially grateful to Trudy Nicholson for producing the beautiful plant and animal drawings that grace this book.

1 Introduction

Long before the concept of biological evolution entered the human mind, people classified diverse forms of life into recognizable categories. Some of the earliest spoken words undoubtedly were names ascribed by primitive peoples to particular types of plants and animals important in their daily lives. Theorists and professional biologists categorized organisms too. For example, in the third century BC the Greek philosopher Aristotle grouped species according to morphological conditions (such as winged versus wingless, and two-legged versus four-legged) that he supposed had been constant since the time of Creation. About twenty centuries later, Carolus Linnaeus – a Swedish botanist and the acknowledged father of biological taxonomy – classified organisms into nested groups (such as genera within families within orders within classes), but still he had no inkling that varied depths of evolutionary kinship might underlie these hierarchical resemblances.

More time would pass before scientists finally began to understand that life evolves, and that historical descent from shared ancestors was responsible for many of the morphological similarities among living (and fossil) species. This epiphany is sometimes mistakenly attributed to Charles Darwin (CD), but several scientists before him in the late 1700s and early 1800s, including Jean-Baptiste Lamarck, Comte de Buffon, and CD's own grandfather Erasmus Darwin, were well aware of the reality of evolutionary descent with modification. What CD "merely" added was the elucidation of natural selection as the primary driving agent of adaptive evolution (this achievement was, of course, one of the most influential in the history of science). The point here, however, is that even before CD, the nested classifications of traditional taxonomy had been interpreted by some systematists as logical reflections of the nested branching structures in evolutionary trees.

The meaning of phylogeny

Evolution has few universal laws, but one unassailable truth is that every organism alive today had at least one parent, who in turn had either one or two parents (depending on whether the lineage was asexual or sexual), and so on extending

back in time. The following imagery may help to convey the incredible temporal durations of these extended hereditary lines. Imagine yourself as the current carrier of a genetic baton that was passed along a multi-generation relay team composed of your direct-line ancestors across the past 200 000 years (*c.* 10 000 generations), beginning when creatures virtually indistinguishable from modern *Homo sapiens* first strolled onto the evolutionary scene. If each successive generation of your predecessors had jogged a quarter mile, your family's cross-country relay squad could have transferred the baton from Los Angeles to New York.

The proto-human lineage is known to have separated from the proto-chimpanzee lineage about five to seven million years ago, so across that longer stretch of geological time your ancestral relay team would have covered a distance equivalent to three times the earth's circumference, or about one-third of the way to the moon. If the evolutionary marathon had been monitored across 40 million years, starting when anthropoid primates first arose, at least two million generations of your ancestors would have come and gone (actually more than that, because monkeys have shorter generation lengths than humans). During that time, your hereditary baton would have been passed a distance of at least half a million miles. This logic can be extended (Dawkins, 2004), ultimately to life's origins nearly four billion years ago. If your extended family lineage had dropped the genetic baton (failed to reproduce) even once during this Olympian marathon, you would not be here. Comparable statements apply to every living creature, each of which is the current embodiment of its own hereditary legacy ultimately stretching back through an *unbroken* chain of descent, with genetic modification, across untold generations.

The word phylogeny (from Greek roots "phyl" meaning tribe or kind, and "geny" meaning origin) refers to the chronicle of life, i.e. to the extended hereditary connections between ancestors and their descendents. Thus, phylogeny can be broadly defined as the evolutionary genetic history of life at any and all temporal scales, ranging from close kinship within and among closely related species to ancient connections between distantly related organisms that last shared common ancestors hundreds of millions of years ago.

In the first 100 years following Darwin, scientists estimated phylogenies for particular taxa by comparing visible organismal phenotypes – e.g. morphological, physiological, or behavioral characteristics – that they could only presume were reflective of underlying genetic relationships. When species were found to share particular phenotypic traits, the usual supposition was that they did so by virtue of shared ancestry. This interpretation was not always correct, however, because some phenotypes arise by convergent evolution. Wings, for example, originated independently in insects, birds, and mammals (plus some other groups,

such as the pterodactyl reptiles of the Mesozoic Era). So, among extant vertebrates (animals with backbones) alone, wing-powered flight evolved at least once in birds and again independently in bats, but this fact becomes apparent only when many other phenotypic features (such as feathers, fur, and pregnancy) are taken into account. Few evolutionary cases are so straightforward, however, and the basic challenge in genuinely intriguing situations is to distinguish phenotypic conditions that provide a valid phylogenetic signal from those that yield mostly phylogenetic noise (i.e. homoplasy).

Following the introduction of various molecular technologies, beginning about 40 years ago, scientists gained powerful new genetic tools to estimate phylogenetic trees for any living species, as connected across any depths in the vast continuum of evolutionary time. This temporal scope is made possible because some DNA sequences have evolved very rapidly, others very slowly, and others at intermediate rates. Fast-evolving sequences are most useful for estimating phylogeny at shallow evolutionary timescales (i.e. for organisms that shared common ancestors within the past few thousands or millions of years), whereas slow-evolving DNA sequences find special utility in estimating phylogeny over much deeper evolutionary timeframes. Few types of molecular trait are themselves free of homoplasy, but when hundreds, thousands, or millions of molecular characters are examined in a given study (as is now routinely the case), empirical experience has indicated that they collectively beam a strong phylogenetic signal.

In 1973, the famous evolutionary geneticist Theodosius Dobzhansky encapsulated a fundamental biological truth in one pithy statement: "Nothing in biology makes sense except in the light of evolution." It is equally true, in turn, that much in evolution makes even more sense in the light of phylogeny. Biological entities are unlike inorganic units (such as gas molecules, or rocks) that can move rather freely in any direction and speed in response to external forces. Instead, the history-laden genetic makeup of organisms directs and constrains each species to a small subset of all imaginable evolutionary trajectories. Each extant species is a current incarnation of an extended lineage whose idiosyncratic genetic past has dictated the present and will also delimit that species' evolutionary scope for the future. Gorillas may dream of flying, but their ponderous bodies of primate ancestry preclude self-powered flight from their foreseeable evolutionary prospects.

Phylogenetic metaphors

Various metaphors can help to capture the general notion of phylogeny. A formerly popular (but invalid) metaphor portrayed evolution as a ladder, the rungs of which held successive forms of life that presumably had climbed higher and

higher toward biological perfection. The lowest rungs were occupied by "lowly" microbes, and atop the highest rung was, of course, *Homo sapiens*. A metaphor with greater legitimacy describes biological lineages as genetic threads stretching back through the ages, and from which the fabric of life has been woven by natural selection and other evolutionary forces including mutation, recombination, and serendipity. As mentioned above, an ineluctable truth is that any lineage alive today extends back generation after generation, ultimately across several billion years to when life originated. Only a minuscule fraction of such hereditary lineages has persisted across the eons to the present; extinction has been the fate of all others. Quite literally, lineages fortunate enough to have survived this epic evolutionary journey have hung on by just a thread.

The eminent paleontologist George Gaylord Simpson invoked another powerful metaphor when he proclaimed: "The stream of heredity makes phylogeny; in a sense, it is phylogeny. Complete genetic analysis would provide the most priceless data for the mapping of this stream." That statement, issued in 1945, was all the more prescient because it came in the "pre-molecular" era, before direct biochemical assays of DNA were available (indeed, even before DNA was firmly documented to be life's hereditary material). Like other biologists of his time, Simpson estimated phylogenies by comparing morphological features among living and fossil species. He none the less appreciated that morphological resemblance is merely a surrogate (and sometimes a rather poor one) for establishing propinquity of descent among the creatures being compared, and that direct genetic analyses eventually would be required. Today, by extracting and comparing DNA sequences from living creatures (and occasionally from well-preserved recent fossils), and by reconstructing phylogenies from those molecular genetic data, scientists can more fully explore both the headwaters and the many forks in the streams that constitute the evolutionary watersheds of life.

Ever since the mid 1800s, however, the most popular metaphor for evolution's pathways has not been ladders, threads, or watersheds, but rather phylogenetic trees (Box 1.1; Fig. 1.1). Under this view, DNA is the sap of heredity that has flowed through the ancient roots, trunks, and branches, and finally into the most recent twigs in various sections of the Tree of Life. The tree analogy for phylogenies is indeed apt (albeit imperfect). Much as twigs and limbs in a botanical tree trace back to successively older forks, so too do living species trace their ancestries back through branched hierarchies of ever-more-ancient phylogenetic nodes. Just as forks in a botanical tree tend to be bifurcate (rather than multi-furcate), most speciational nodes in a phylogenetic tree are dichotomous. Much as a real tree fosters new growth primarily from its growing tips and buds, biodiversity at any point in evolutionary time propagates exclusively from then-extant species.

Box 1.1 Basic definitions regarding phylogenetic trees

See Fig. 1.1 for examples; see also Box A1 in the Appendix, and the Glossary, for additional relevant terms and concepts.

(a) *phylogenetic tree (phylogeny)*: a graphical representation of evolutionary genetic history.
(b) *phylogenetic network*: an unrooted phylogeny (e.g. diagram I in Fig. 1.1).
(c) *root*: the most basal branch (pre-dating the earliest node) in a phylogenetic tree (the thick line at the left in diagrams II and III of Fig. 1.1).
(d) *branch*: an extended ancestral–descendent lineage between nodes in a phylogenetic tree.
(e) *interior node*: a branching point inside a phylogenetic tree (i.e. an ancestral point from which two or more branches stem, or, from the perspective of the present, an ancestral point to which any specified set of extant lineages coalesces). In Fig. 1.1, interior nodes are indicated by black dots labeled with the lower-case letters g–k. In any phylogenetic network, interior nodes can be thought of as ball-and-socket joints around which branches can be freely rotated without materially affecting network structure. Thus, angles between branches have no meaning. Similarly, branches can be rotated around interior nodes in a rooted phylogenetic tree, but only in the vertical plane.
(f) *exterior node*: an outer tip on a phylogenetic network or tree, usually representing an extant species (e.g. A–F in the diagrams of Fig. 1.1).
(g) *operational taxonomic units (OTUs)*: the biological entities (e.g. DNA sequences, individuals, populations, species, or higher taxa) analyzed and depicted in a particular phylogenetic tree (again, A–F in Fig. 1.1).
(h) *anagenesis*: genetic change within a lineage (along one branch of a phylogenetic tree) through time.
(i) *cladogenesis*: the splitting or bifurcation of branches in a phylogenetic tree (normally equated with speciation).
(j) *cladogram*: a representation of cladistic relationships, i.e. of a phylogenetic tree's hierarchical branching structure (but otherwise implying nothing about branch lengths).
(k) *phylogram*: a representation of a phylogenetic tree that includes information on branch lengths in addition to cladistic (branching) relationships.
(l) *phenogram*: a tree-like depiction that summarizes overall phenetic (not necessarily phylogenetic) relationships among a set of organisms.
(m) *gene tree*: a graphical representation of the evolutionary history of a particular genetic locus (as opposed to the composite organismal phylogeny of which any gene tree is only a small component).

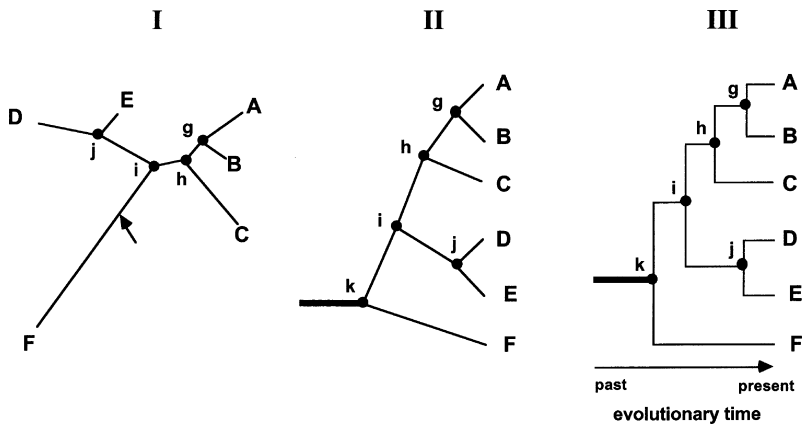

Figure 1.1. Phylogenetic trees. Diagrams I–III illustrate alternative but essentially equivalent representations of evolutionary relationships among six hypothetical extant species (A–F). Diagram I is an unrooted tree (phylogenetic network), whereas phylogenetic trees II and III are rooted (at the position of the arrow in diagram I). See Box 1.1 for additional definitions and descriptions.

One shortcoming of the tree metaphor, however, is that real trees have a large trunk and successively smaller branches and twigs, whereas hereditary routes in phylogenetic trees have no distinct tendency to decrease (or increase) in diameter across evolutionary time. In a phylogenetic tree, what split at each node are particular biological species, rather than composite collections of independent species. For example, birds did not evolve from reptiles collectively; rather, one or a few related reptilian species in the Mesozoic Era gave rise to particular proto-avian species from which all other birds eventually descended. For this reason, all phylogenetic trees depicted in this book will be drawn as stick-like diagrams with more or less uniform branch width. In addition, to make labeling easier, nearly all phylogenetic trees presented here will be rotated through 90° relative to an upright real tree, such that the right terminus of each diagram indicates the present time and successive nodes to the left reflect progressively older dates in the evolutionary past.

Charles Darwin included only one figure in his 1859 masterpiece *The Origin of Species*. It was of a phylogenetic tree (albeit an unattractive rendition). However, Ernst Haeckel (a German philosopher and evolutionary biologist) did far more to make an iconography of the tree metaphor by gracing his 1866 book – *Generelle Morphologie der Organismen* – with lovely arbor diagrams, one of which is shown here in Fig. 1.2. Haeckel drew his trees as literal metaphors, complete with bark and gnarled branches. There is, however, a serious shortcoming (apart from the branch-width issue mentioned above) in the style of Haeckel's depictions: namely,

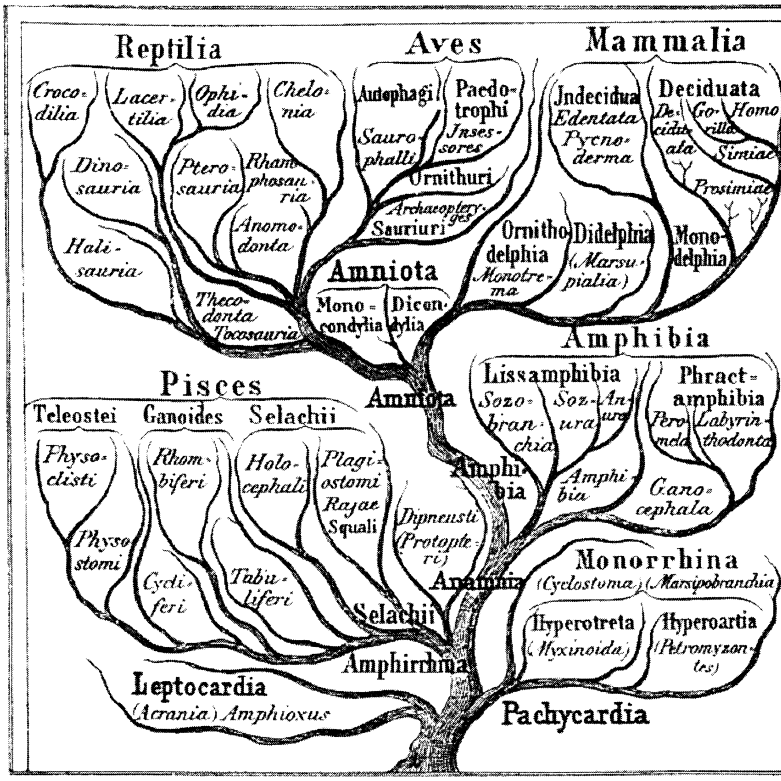

Figure 1.2. Example of a phylogenetic tree from Haeckel's (1866) *Generelle Morphologie der Organismen.*

they convey an impression that some living species (such as birds and mammals) are higher in the Tree of Life than others (such as fishes and amphibians), when in fact all lineages leading to extant forms of life have maintained continuous genetic ancestries that trace back ultimately to life's geneses. Thus, if height above the ground in Haeckel's trees is taken to imply the duration of evolutionary existence, the depictions are misleading, because by this criterion all extant branch tips are, in truth, equally high. This is another reason why most of the phylogenetic trees depicted in this book have right-justified branch tips.

Most of our scientific understandings about biology can be improved by implicit or explicit reference to well-grounded phylogenetic trees. For example, the basic challenge in the science of systematics is to describe various portions in the Tree of Life, i.e. to reconstruct the temporal order of forks (speciation events), to measure branch lengths (the amount of genetic change along each branch through time), and to estimate how many buds (distinct species and populations) currently exist

from which any future growth might ensue. A primary aim of the conservation sciences is to promote the survival and the potential for diversification of the outermost tips in the Tree of Life. This task is daunting because a burgeoning human population, through its direct and indirect impacts on the environment, threatens to prune if not defoliate much of the Tree's luxuriant canopy. Societies must find better ways to identify, characterize, and protect the vigorous as well as the most tender of extant shoots so that, in this latest instant of geological time, humankind does not terminate what nature had germinated and assiduously propagated across the eons. And finally, in the sciences of ecology, paleontology, ethology, natural history, and evolutionary biology, a fundamental challenge is to understand the historical origins of species and their diverse phenotypes. As I attest via this book, all of these tasks demand an appreciation of phylogeny.

Molecular appraisals of phylogeny

In 1963, a biochemist in Chicago reported a discovery that would prove to be a major conceptual step forward in the field of phylogenetic biology. By compiling and scrutinizing published information on cytochrome *c* – a protein involved in cellular energy metabolism, and consisting of a molecular string of 104 amino acids (the building blocks of proteins) – Emanuel Margoliash (1963) found that these molecules differed in structure, to varying degrees, among human, pig, horse, rabbit, chicken, tuna, and baker's yeast. For example, the cytochromes *c* of horse and pig differed at three amino acid positions along the molecular string, whereas those of horse and tuna differed at 19 amino acid positions and horse and yeast differed at 44 such sites. These differences in amino acid sequence reflect the evolutionary accumulation of underlying mutations in the DNA molecules (i.e. the genes) that encode cytochrome *c*. Margoliash concluded that "The extent of variation among cytochromes *c* is compatible with the known phylogenetic relations of species. Relatively closely related species show few differences . . . phylogenetically distant species exhibit wider dissimilarities."

From a phylogenetic perspective, there is nothing special about cytochrome *c*. It is merely one of many thousands of cellular proteins, each encoded by a different functional gene. The genes themselves consist of long strings of four types of molecular subunits – the nucleotides adenine (A), thymine (T), cytosine (C), and guanine (G) – that make up not only protein-coding DNA sequences but also vast stretches of non-coding DNA. The collective lengths of these nucleotide strings are astounding. For example, each copy of the human genome (a full suite of DNA in each of our cells) consists of more than three billion pairs of nucleotides wedded into strands that give DNA its double helical structure. The genomes of most other vertebrates are roughly similar in size, and those of various species of invertebrate

animals, fungi, and plants range in length from about 10 million to more than 200 billion nucleotide pairs.

Margoliash's findings provided one of the first clear indications that DNA sequences sampled from organismal genomes gradually accumulate specifiable molecular differences during the course of evolution, and that "the extent of variation of the primary structure . . . may give rough approximations of the time elapsed since the lines of evolution leading to any two species diverged." We now know that, during their passage across large numbers of successive generations, DNA molecules (and hence any protein molecules they may encode) often tend to evolve in clock-like fashion. Although molecular clocks are far from metronomic – they tend to tick at somewhat different rates depending on the lineage and on the specific type of DNA sequence examined (see Box 1.2) – they none the less can be informative about approximate nodal dates in evolutionary trees. Furthermore, some methods for estimating phylogenetic trees depend hardly at all on a clock-like behavior (see the Appendix). For example, by considering the evolutionary chains of mutational events required to convert one DNA sequence into another, branching topologies of evolutionary trees can often be recovered even when precise evolutionary dates cannot be attached to particular internal nodes. The bottom line is that, when researchers sample and compare long molecular passages from organisms' genomic archives, they can deduce how various living species have been connected to one another in their near and distant evolutionary pasts.

Box 1.2 DNA sequences for molecular phylogenetics

Many different types of DNA sequence are employed to estimate organismal phylogeny, the choice in each instance dictated by the general evolutionary timeframe under investigation and also by numerous technical considerations. The following are introductory comments about some of the gene sequences widely used in comparative phylogenetics.

Cytoplasmic genomes

These are relatively small suites of DNA housed inside organelles within the cytoplasm of eukaryotes (organisms whose cells have a distinct membrane-bound nucleus). The two primary cytoplasmic genomes are mtDNA in the mitochondria of animals and plants, and cpDNA in the chloroplasts of plants.

In animals, mtDNA is usually a closed-circular molecule about 16 000 to 20 000 nucleotide pairs long. It typically consists of 37 functional genes: two ribosomal (r) RNA loci, 22 transfer (t) RNA loci, and 13 structural genes specifying polypeptides (protein subunits) involved in cellular energy production.

The molecule tends to evolve quite rapidly overall, thus making it suitable for phylogenetic appraisals at micro-evolutionary scales (e.g. of conspecific populations), and also across meso-evolutionary timeframes (i.e. for species that separated up to scores of millions of years ago). Different mtDNA loci evolve at quite different rates, however, with some (such as the control region) diverging very rapidly and others (such as the rRNA loci) evolving far more slowly. Thus, with appropriate choice of mtDNA sequences, phylogenetic studies can be tailored to varying evolutionary timescales.

Full-length mtDNA molecules in plants are much larger (200 000 to more than 2 000 000 nucleotide pairs long, depending on the species) and for various technical reasons have not proved particularly useful for phylogenetic reconstructions. By contrast, plant cpDNAs offer powerful phylogenetic markers. These closed-circular molecules, ranging from about 120 000 to 220 000 nucleotide pairs long, generally evolve at a leisurely pace, so their sequences tend to be especially suitable for phylogenetic estimates among plant genera, families, and taxonomic orders.

Nuclear genomes

In any eukaryotic cell, most of the tremendous variety of DNA sequences is housed within the nucleus. For example, each complete set (i.e. haploid copy) of the human nuclear genome consists of more than three billion nucleotide pairs arranged along 23 chromosomes. The nuclear genome of a typical eukaryotic species (humans included) contains about 25 000 protein-coding genes, one or a handful of which are normally sequenced from multiple species in a conventional molecular phylogenetic analysis. Useful phylogenetic information can also be recovered from other nuclear regions such as rRNA loci, regulatory domains flanking structural genes, or particular subsets of non-coding sequences (often highly repetitive) that actually make up the great majority of nuclear DNA in most species.

Combined information

Typical molecular phylogenetic analyses conducted to date have involved DNA sequences from several nuclear or cytoplasmic genes (or both) totaling about one thousand to several thousand nucleotide pairs per specimen. With continuing improvements in DNA sequencing technologies, the standards are quickly being raised. For example, it has become almost routine in recent years for phylogeneticists to sequence entire 16 kilobase mtDNA genomes from the animals they survey.

Representative genomes of approximately 1000 species (humans included) have been fully sequenced in recent years, and substantial amounts of DNA sequence data are rapidly accumulating for many thousands more. Today, scientists routinely read these genetic scriptures to reconstruct the histories of life. As Margoliash correctly presaged in 1963, molecular details in the genomic registries provide "a faithful recorder of the unit events of evolution." Scientists are no longer content, however, to draw crude phylogenetic sketches for a miscellany of distant organisms such as human, rabbit, chicken, tuna, and yeast. Instead, they now use extensive molecular data to paint detailed evolutionary pictures for hundreds of species of mammals, birds, reptiles, amphibians, and fishes, plus all sorts of invertebrate animals, fungi, plants, and microbes. In the past decade or two, molecular phylogenetics has grown into one of the most active areas in all of biological research. Perhaps 10 million or more species currently inhabit the planet, so reconstruction of the full Tree of Life will require a huge scientific effort. None the less, DNA sequences recovered to date already permit solid phylogenetic estimates for many taxa, and the molecular phylogenies in turn can be employed to chart the evolutionary courses of organismal phenotypes (morphological, behavioral, and so on). This book will delve into some of the most evocative and sometimes controversial of these phylogenetic mapping exercises to date.

Comparative phylogenetics

Evolutionary biologists routinely employ "comparative phylogenetic" methods, by which can be meant many things. In a catholic sense, any phylogenetic procedure is comparative if it involves more than one gene, more than one phenotype, more than one taxonomic group, or any combination of the above. For example, it is perfectly permissible and often highly informative to compare a phylogenetic estimate based on one set of phenotypes with that based on another set of phenotypes, or to compare the phylogenies of two or more taxonomic groups against a geographic backdrop (for example) that may have influenced their evolutionary histories. In other words, the basic idea of comparative phylogenetics is to compare and contrast historical evolutionary patterns across multiple types of characters or taxa.

The molecular revolution in evolutionary biology, which began in the second half of the twentieth century, ushered in another powerful way to conduct comparative phylogenetics. Specifically, it afforded access to a potentially huge set of DNA-level and protein-level characters that could be employed as a basis for phylogenetic comparisons with organism-level traits. This book will focus on how molecular estimates of phylogeny have informed our understanding of the ways and means by which organismal phenotypes evolve. However, I hope not to be interpreted

as chauvinistic with regard to molecular approaches. The fact is that systematists practiced comparative phylogenetics, widely and fruitfully, long before molecular genetic techniques became available. The comparisons then were based on visible phenotypes and other traditional systematic characters, many of which are readily accessible and hugely informative in their own right about phylogenetic relationships. Molecular data have added another comparative layer to phylogenetic practices, thereby enriching a field that already had a long and productive scientific tradition.

As applied here in a narrower sense, the term comparative phylogenetics will mean any application of DNA-based phylogenies with the intent of revealing the evolutionary histories of organismal phenotypes. In this explicit subcategory of comparative phylogenetic analysis, four steps are normally entailed: (i) DNA-sequencing methods or other laboratory techniques are used to gather extensive molecular data from homologous genes in living species; (ii) based on that genetic data, a phylogeny for those species is estimated by using appropriate tree-building algorithms; (iii) particular phenotypic characters showing variation (such as winged versus wingless) among the species of interest are examined to establish their present-day taxonomic distributions; and (iv) the phylogenetic histories of those phenotypes are provisionally reconstructed by plotting their inferred ancestral states and evolutionary interconversions along various branches of a molecular tree. The first three steps can be thought of as background and the fourth step as the crux of the process of "phylogenetic character mapping" (see below).

A vast technical literature exists on molecular methodologies and phylogenetic algorithms underlying steps (i) and (ii). A cursory introduction is provided in Box 1.3, but readers interested in further details are directed elsewhere (see the references listed). Fortunately, for current purposes, a thorough understanding of molecular techniques and phylogenetic reconstruction methods is not a prerequisite for appreciating the biological discoveries about nature that are the focal points of this book.

Step (iii) normally involves relatively straightforward description, except that questions may arise about how to define and characterize organismal phenotypes. For example, these phenotypes may be alternative qualitative states of composite structures (such as wing-presence versus wing-absence) or behaviors (e.g. flighted versus flightless), or more narrowly defined characters (such as wings made of skin flaps versus those made of feathers and with interior bones; or flapping versus gliding flight). Always, evolutionary interpretations must be adjusted accordingly. For example, the broad attribute "flight" is clearly polyphyletic (i.e. arose on multiple evolutionary occasions) in animals, whereas more specific characteristics often associated with flight (such as feathers in birds, echolocation in bats, or presence

Box 1.3 Molecular methods and phylogenetic algorithms

Steps (i) and (ii) in comparative molecular phylogenetics (see text) entail the acquisition and phylogenetic analysis of molecular data. These are vast topics, well beyond the scope of this book, so only a brief introduction and some key references for interested readers are provided here.

Molecular methods

Many types of laboratory assay have been developed for retrieving molecular information from organismal genomes. Most of the earlier methods accessed DNA sequences indirectly, for example through assays of proteins, or via quantitative biochemical techniques such as DNA–DNA hybridization (a technique that yields numerical estimates of genetic divergence by examining the thermostabilities of nucleotide sequences). These and other tried-and-true molecular methods have been employed widely to generate phylogenetic trees as evolutionary backdrop for phylogenetic character mapping (PCM).

One of the most powerful of the modern molecular techniques – DNA sequencing – directly elucidates the precise sequences of nucleotides along specified stretches of DNA. In the past decade, refinements in laboratory methods have permitted scientists to generate large amounts of DNA sequence data, and thereby have made DNA sequencing today's most popular approach in comparative phylogenetics.

Recommended reading: Avise, 2002 (beginner level); Avise, 2004 (intermediate); Baker, 2000 (intermediate); Hillis *et al.*, 1996 (advanced).

Phylogenetic methods

Many data-analysis procedures (usually implemented in powerful computer programs) are available for reconstructing phylogenetic trees from molecular data. However, all such methods can be characterized as beginning with either: (a) numerical estimates of genetic distances among taxa (as obtained, for example, from DNA–DNA hybridization, or from tallies of nucleotide differences obtained directly by DNA sequencing); or (b) the raw character states themselves (such as specified nucleotides at many successive positions along particular stretches of DNA). In the former method, pairwise values in a matrix of genetic distances between species are "clustered" to yield an estimate of phylogeny according to user-specified algorithms (there are several available options). In the latter method, the qualitative data in DNA sequences from multiple taxa are analyzed directly to yield phylogenetic estimates based on particular assumptions or models (again there are many available options)

about the nature of evolutionary interconversions among those character states.

Even with computer assistance (such as by the software PAUP (Swofford, 2000)), the search for the "best" tree is daunting when more than a few species are being compared, in part because the number of possible phylogenetic arrangements among multiple taxa is astronomical. For example, for merely 10 species the potential number of different bifurcating tree structures is more than 30 million, and for 20 taxa that number becomes about 8.2×10^{21}! From among such vast numbers of candidate trees, the objective is to identify phylogenetic arrangements that closely approximate the true evolutionary history of the taxa examined. In effect, phylogenetic algorithms in computer programs often search among possible trees for those that best comply with some user-specified evolutionary model or optimality criterion. For example, parsimony approaches (which themselves have several versions) generally operate under the assumption that the preferred tree(s) are those with the shortest total branch lengths (i.e. the fewest possible evolutionary interconversions among character states) consistent with the empirical data. (However, it should also be remembered that evolution does not always proceed along most-parsimonious routes.)

In recent years, maximum likelihood (ML) and Bayesian methods have also become popular ways to analyze molecular data and to statistically test competing phylogenetic hypotheses (see the reviews by Huelsenbeck and Rannala, 1997; Holder and Lewis, 2003). These conceptually related approaches entail computer-based explorations of tree structures (and their associated probabilities) that best explain the underlying data under specifiable models of molecular evolution. The development of fast computer software programs such as TREE-PUZZLE for ML (Strimmer and von Haeseler, 1996) and MRBAYES for Bayesian approaches (Huelsenbeck, 2000) has greatly facilitated the implementation of these newer phylogenetic methodologies.

Recommended reading: Hall, 2004 (beginner level); Avise, 2004 (beginner); Nei and Kumar, 2000 (intermediate); Hillis et al., 1996 (intermediate); Li, 1997 (intermediate); Felsenstein, 2004 (advanced).

of compound eyes in some insects) might each have arisen once or only a few times during evolution. Many other phenotypes (such as feather density or the number of facets in a compound eye) may vary more or less continuously, rather than qualitatively, among an array of taxa; such quantitative characteristics with numerous states often pose some of the greatest challenges for proper phylogenetic interpretation.

Throughout this book, I will use the words "characters," "traits," "features," "conditions," and "attributes" more or less interchangeably to mean multiple

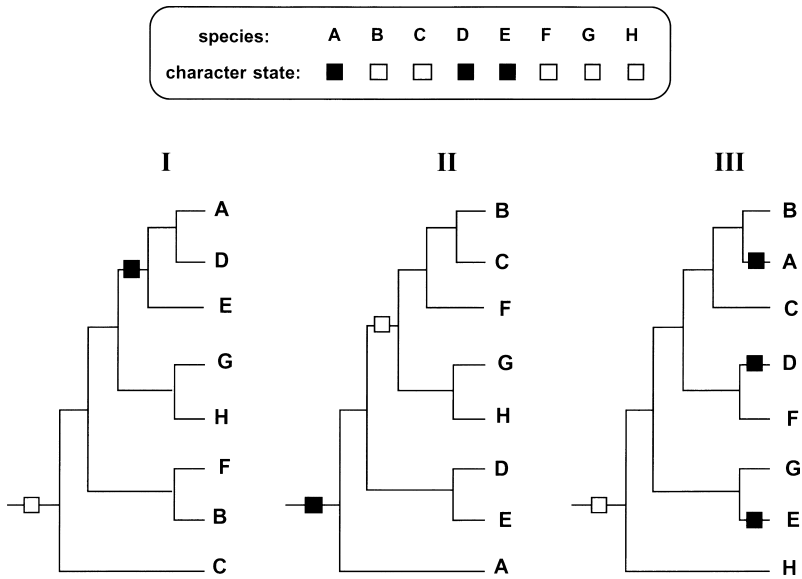

Figure 1.3. Introduction to the basic conceptual approach of PCM. Shown across the top are eight hypothetical species (A–H) displaying one or the other of two character states (white squares or black squares) of a particular phenotype. How these character states most likely evolved can be informed by knowledge about these species' evolutionary relationships (as can be estimated using molecular genetic data). For example, if species A–H prove to be phylogenetically related as shown in diagram I, then white-square was probably the ancestral condition for the entire group, and black-square is a shared derived condition (i.e. a synapomorphy; see Box A1) for the ADE clade. However, if the species are phylogenetically allied as shown in diagram II, then black-square was probably the original ancestral condition and white-square is a shared derived state for clade BCFGH. Many other outcomes are also possible. For example, if the true phylogeny for species A–H is as shown in III, then white-square was probably the ancestral condition from which black-square evolved independently on three separate occasions.

states of any specified phenotypes. In this generic usage, organismal characteristics to be phylogenetically mapped may encompass phenotypic descriptions of any sort (e.g. qualitative or quantitative) and at any indicated level of inclusiveness (from broad trait descriptions to those that are highly detailed). This means that biological interpretations of evolutionary outcomes will vary according to the particular phenotypes examined and questions addressed in each phylogenetic analysis.

Phylogenetic character mapping

Step (iv) mentioned above, the primary component of comparative phylogenetics treated in this book, has sometimes been referred to as evolutionary trait analysis,

comparative trait charting, or phylogenetic character mapping (henceforth PCM). Under the PCM approach, alternative states of particular phenotypic characters are matched with their associated species on an independently established phylogeny, the purpose being to reveal the evolutionary origins of those phenotypes and their probable patterns of historical interconversion. A primer to the basic methodological concept of PCM is presented in Fig. 1.3, and a somewhat fuller introduction (for the uninitiated) is provided in the Appendix. Some readers may wish to examine the Appendix first, as further technical background, before proceeding to the empirical studies presented in Chapters 2–7.

Many challenging conceptual and operational issues surround comparative phylogenetics (see Box 1.4); these too are discussed at length in an extensive scientific

Box 1.4 Acknowledged limitations of comparative molecular phylogenetics

For most of the case studies presented in this book, the phylogenetic reconstructions and biological conclusions faithfully reflect those of the original authors. Thus, I have assumed that the molecular phylogenies and the PCM reconstructions were basically correct as published. This may not invariably be true, of course. Indeed, the history of comparative phylogenetics would seem to suggest that substantial fractions of published interpretations are challenged to varying degrees, sooner or later, by at least some independent researchers. The scientific sources of the resulting phylogenetic controversies can be many. Listed below are examples of hard questions that critical readers should ask before accepting the face-value conclusions from any PCM analysis.

Does the molecular phylogeny itself correctly reflect the species phylogeny?
e.g. . . .

What types of molecular genetic assay were conducted and were they reliable?

How many unlinked genes were assayed, and how long were their sequences? [When the volume of genetic data is small, gene trees can be poor or misleading indicators of overall or composite relationships in a species tree (see, for example, Rokas *et al.*, 2003).]

Were evolutionary convergences or reversals of character states (i.e. homoplasy) likely?

Were assumptions underlying the phylogenetic analyses properly suited to the category of molecular data gathered?