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).

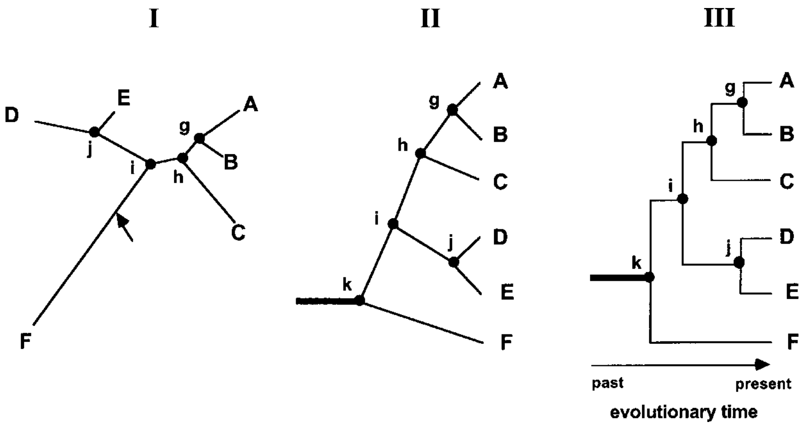

Figure 1.1. Phylogenetic trees. Diagrams I–III illustrate alternative but essentially equivalent representations of evolutionary relationships among six hypothetical extant species (A–F). Diagram I is an unrooted tree (phylogenetic network), whereas phylogenetic trees II and III are rooted (at the position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

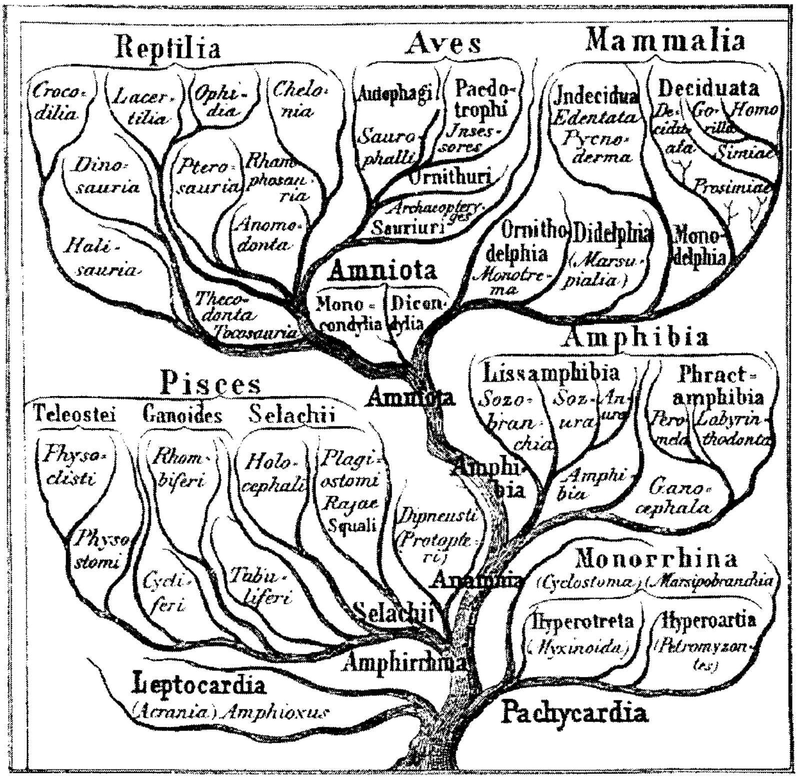

Figure 1.2. Example of a phylogenetic tree from Haeckel's (1866) *Generelle Morphologie der Organismen*.

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.