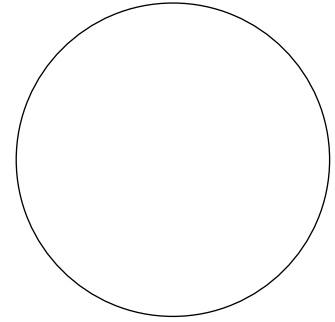


CHAPTER 13

Phylogeny and systematics



Phylogeny refers to the history of a species, to its relationships to other species (in Greek *phyl-* refers to tribe; *gen-* refers to origin or descent). Systematics refers to the methods used to discover that history (in Greek *systematos* refers to a complex whole put together). At the largest scale, the group studied is all of life on Earth, and the goal is to discover the entire pattern of relationships of all things of whose existence we have some evidence. As Darwin put it (1859):

It is a truly wonderful fact—the wonder of which we are apt to overlook from familiarity—that all animals and all plants throughout all time and space should be related to each other in groups . . . The several subordinate groups in any class cannot be ranked in a single file, but seem to be clustered around points, and these around other points. If species had been independently created, no explanation would have been possible of this kind of classification . . . The affinities of all the beings of the same class have sometimes been represented by a great tree. I believe this simile largely speaks the truth . . . The green and budding twigs may represent existing species; and those produced during former years may represent the long succession of extinct species . . . As buds give rise by growth to fresh buds, and these, if vigorous, branch out and overtop on all sides many a feebler branch, so by generation I believe it has been with the great Tree of Life, which fills with its dead and broken branches the crust of the earth, and covers the surface with ever-branching and beautiful ramifications.

So important was this concept to Darwin that he expressed it in the only illustration in *The Origin of Species* (Figure 13.1).

The Tree of Life has three main branches: Bacteria, Archaea, and Eukaryota

Today the methods of molecular systematics, introduced below, have allowed us to reconstruct the major features of the Tree of Life. Any complete branch of the Tree of Life, large or small, recent or ancient, is called a **clade**—a natural group of related organisms that all share a most recent common ancestor. At the largest scale, the Tree of Life is primarily a tree of bacteria with three clades—the Bacteria, the Archaea, and the Eukaryota—rooted at about 3.7 billion years

● KEY CONCEPT

All living things are related. By inferring their relationships, we can reconstruct the history of life on this planet.

Clade: A natural group of related species containing all descendants of the most recent common ancestor of the group.

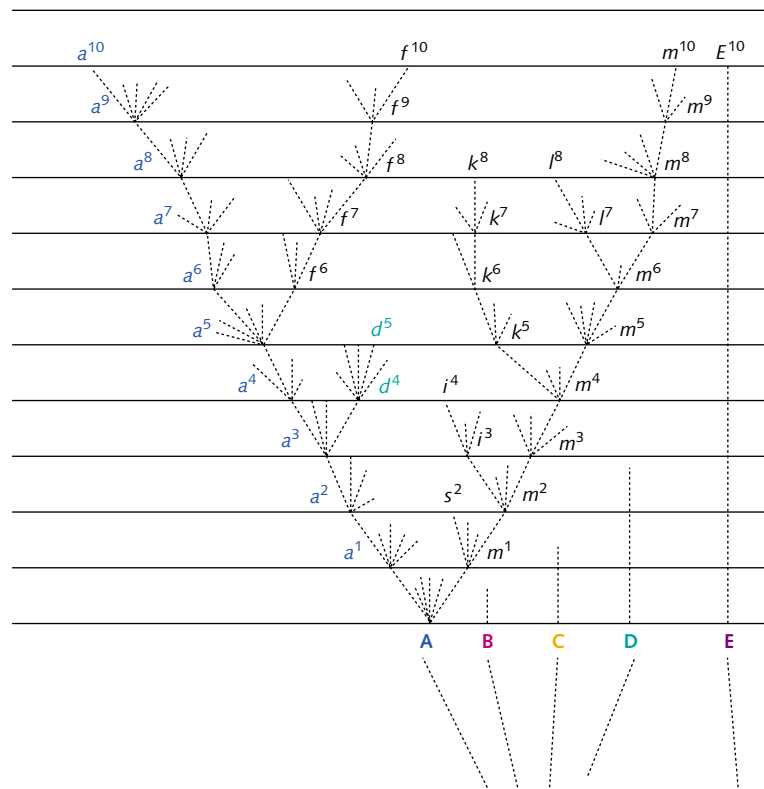


Figure 13.1 Darwin's picture of a phylogenetic tree. Time runs from bottom to top, a convention preserved to this day. Of the five species present at the start—A, B, C, D, and E—only A and E have left living descendants. The others went extinct in the order B, C, D. Of the many descendants of A, only a^{10} , f^{10} , and m^{10} are living in the 10th time interval. (From Darwin 1859.)

ago (Figure 13.2). The three main divisions of life are well supported; the position of the root of the tree, marking the origin of life, is difficult to infer. There is some support from ancient gene duplications for placing the root on the branch leading to the Bacteria, which would make the Archaea and the Eukaryotes sister groups. At this scale the plants, animals, and fungi are small branches that you can locate to the left of the Eukaryote stem.

The Bacteria are common prokaryotes living in virtually all environments; they include the human gut commensal *Escherichia coli*, soil bacteria like *Bacillus subtilis*, as well as pathogens like *Salmonella*, *Staphylococcus*, and *Helicobacter*. The Archaea were discovered relatively recently; like Bacteria, they are small, single-celled prokaryotes. Some inhabit extreme environments with high temperatures or high salt concentrations; others inhabit normal seawater. The Eukaryotes are derived from a symbiotic event in which the proto-eukaryote ancestor, which probably had already evolved a nucleus and a cytoskeleton, acquired a proteobacterium—the group of bacteria that contains *Escherichia* and *Agrobacterium*—that evolved into the mitochondrion. (Look for *mitochondrion* in the lower right of Figure 13.2.) Later the chloroplast, which originated as a cyanobacterium, was independently acquired at least three times: in the lineages leading to the green algae and higher plants, to the red algae, and to an obscure group called the glaucocystophytes. (Look for

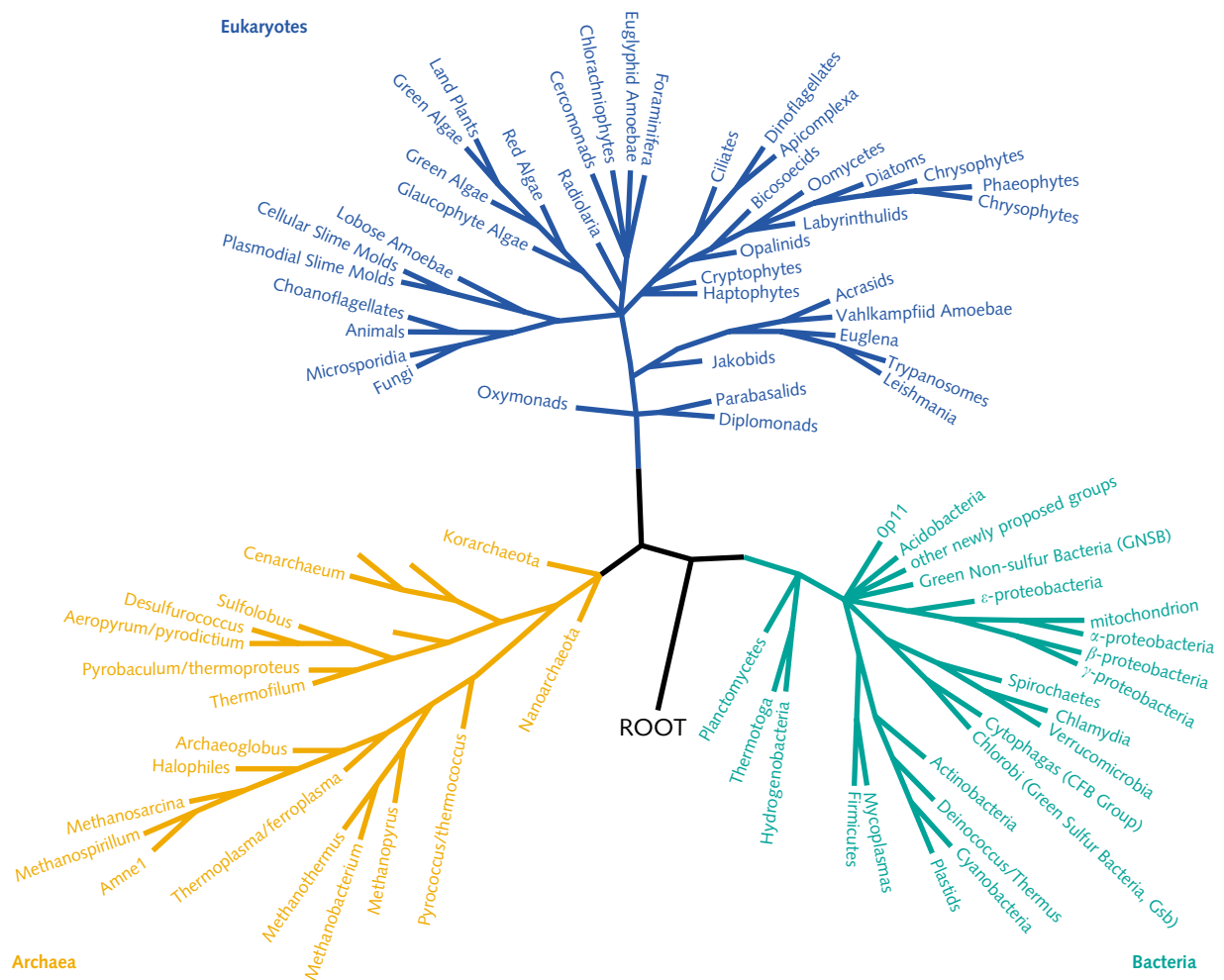


Figure 13.2 The Tree of Life on planet Earth begins about 3.7 billion years ago. There are three major branches—the Bacteria (e.g. *Bacillus*), the Archaea (e.g. *Methanobacterium*), and the Eukaryota (e.g. fungi). At this scale the plants, animals, and fungi represent minor, late radiations. (From Baldauf 2003.)

plastids in the bottom right of Figure 13.2.) Thus all of the ability of eukaryote cells to respire and to photosynthesize was acquired from bacterial symbionts.

Multicellular organisms form three main groups: fungi, plants, and animals.

In the **radiation** of multicellular organisms, three of the major modern groups—plants, animals and fungi—share a common ancestor at about 800–1000 million years ago (Figure 13.3). The times at which various groups originated are given by the concentric circles, one for each 100 million years from 800 million

Radiation: The diversification and divergence of species within a group with a single common ancestor (a clade).

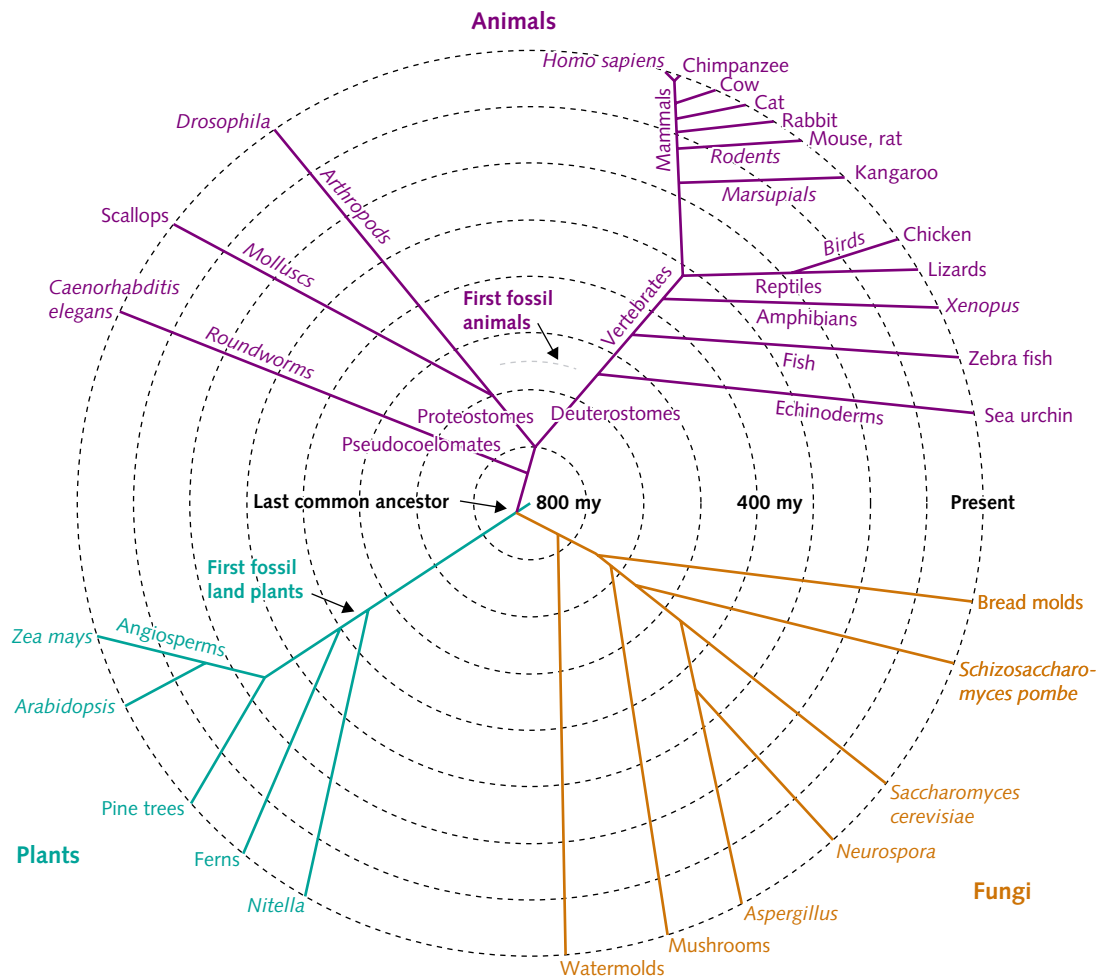


Figure 13.3 The phylogenetic tree of multicellular life, rooted at about 800 million years, with three main branches: plants, fungi, and animals. (Courtesy of T.D. Pollard.)

years ago to the present. Fossils of animals with skeletons are first available in the Cambrian era, about 550 million years ago. They include arthropods, mollusks, echinoderms, and chordates. The first fossil land plants are known from about 440 million years ago, and the angiosperms diverged from the conifers about 220 million years ago. Several major groups—the roundworms, mollusks, and arthropods—appear to have originated before the Cambrian era, as did the line leading to the echinoderms and vertebrates. The branch leading to the mammals split off from the other ‘reptiles’ about 300 million years ago, while the branch leading to the birds split off from the lizards about 200 million years ago. Some of these dates, estimated from molecular divergence, are older than those estimated from fossil evidence.

The Tree of Life is not obvious; it must be discovered

Without phylogenetic trees—pictures like Figures 13.2 and 13.3—we cannot make much sense of the pattern of life. Related organisms are similar in many important ways, and members of one group differ from members of other groups in just as many important ways. Thus the Tree of Life is essential to understanding much of biology, but it is not given to us: it must be discovered. As Darwin (1859) put it:

Our classifications will come to be, as far as they can be so made, genealogies; and will then truly give what may be called the plan of creation. The rules for classifying will no doubt become simpler when we have a definite object in view. We possess no pedigree or armorial bearings; and we have to discover and trace the many diverging lines of descent in our natural genealogies, by characters of any kind which have been long inherited.

Early evolutionary biologists could only use morphological **characters** to work out relationships. Today those characters that ‘have been long inherited’ include DNA sequences, which systematic biologists supplement with data of many other sorts, mainly morphological and developmental. The use of molecular techniques has proven to be extremely powerful in systematics, for it has provided methods to rapidly measure and properly compare a vast number of characters—long DNA sequences from many species—that yield estimates of relationship of unprecedented accuracy reaching in some cases far back into geological time.

Before considering the methods of systematics, we look at some striking recent discoveries powered by these new methods that modified or overturned previous ideas of relationship or otherwise provided unexpected insights.

Character: *A trait that varies among taxa and that in any given taxon takes one out of a set of two or more different states.*

Molecular systematics has yielded surprising insights

Relatives of jellyfish evolved into intracellular parasites

Myxozoa or myxosporidians are single-celled parasites of invertebrates and fishes that cause serious damage. After penetrating the skin they have a complex fungus-like life cycle in the cells of their hosts, where they eat cytoplasm, reproduce sexually, and form spores. The spores resemble the stinging cells of coelenterates, such as jellyfish, not fungal spores (Figure 13.4). Myxozoa were traditionally classified with another group of single-celled parasites, the microsporidia, which are close relatives of fungi, and in 1989 they were placed in a major group of their own. Now molecular systematics identifies them as highly modified cnidarians, placing them within the group that contains the jellyfish, corals, and sea anemones.

● KEY CONCEPT

Much of the history of life that has been obscured by the evolution of morphology can be recovered from DNA sequences.

Figure 13.4 The life cycles of myxozoans (left), which turn out to be highly modified relatives of jellyfish (right). The relationships of this problematic group of fungus-like parasites were resolved by molecular systematics. (By Dafila K. Scott.)

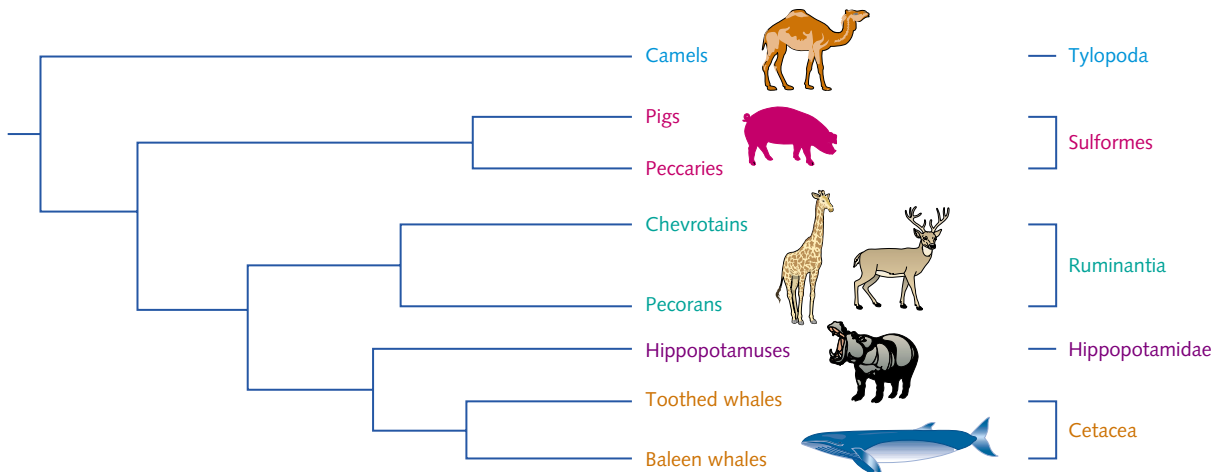
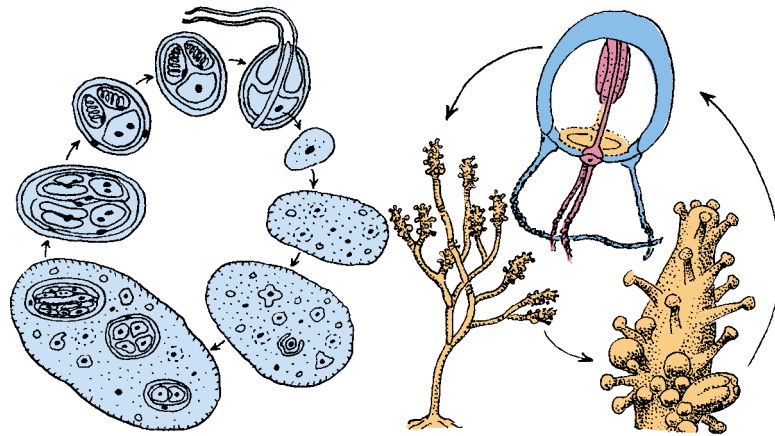


Figure 13.5 The closest relatives of whales are hippopotamuses, and whales are clearly ungulates—highly modified for life in the sea. (Reproduced from Nikaido et al. (1999) Phylogenetic relationships among cetartiodactyls based on insertions of short and long interspersed elements. *Proceedings of the National Academy of Sciences of the United States of America* 96, 10261–6. National Academy of Sciences, USA.)

Whales are ungulates; their closest relatives are hippopotamuses

Whales are mammals that re-entered the sea after a long period on land. Because their highly modified morphology makes their relationship to other mammals obscure, they are prime candidates for molecular detective work. Recent systematic analysis of DNA sequences has revealed that whales are ungulates—not carnivores related to seals and otters—and relatively recently derived ungulates at that. Their closest relatives are hippopotamuses, and their next-closest relatives are deer and antelope (Figure 13.5).

Giant pandas and lesser pandas are not each other's closest relatives

The giant panda, a bamboo-eating specialist now found only in western Szechwan, looks like a bear, but its genital anatomy and vocalizations are unlike any bear, and it has an extra digit, derived from a wrist bone, found in no other animal, that it uses to grasp bamboo shoots. Its closest relative was long thought to be the lesser panda, which also lives in Asia, eats bamboo, and whose molar teeth resemble those of the giant panda. The lesser panda has a long, ringed tail and resembles a raccoon or coatimundi. Molecular systematics have now shown that the giant panda is most closely related to the bears, and the lesser panda now believed to be more closely related to the mustelids (the weasels and their relatives) and to the procyonids (the raccoons and their relatives) than it is to the bears. The similarity of their molar teeth, adapted to bamboo-feeding in both species, misled us about their relationships, for their teeth resembled each other because natural selection had shaped them to similar tasks, not because that morphology had been inherited from a common ancestor (Figure 13.6).

Who transmitted HIV to a rape victim? Systematics can help to identify criminals

The methods of systematics are not limited to basic questions about the Tree of Life. They can be applied to very practical problems. Suppose you are a detective asked by the police to identify which of two suspects transmitted the HIV virus to a rape victim who developed AIDS after the rape was committed. You



Figure 13.6 The giant panda and the lesser panda, two mammals that both eat bamboo and live in Asia. Molecular systematics have shown that the giant panda is related to the bears, whereas the lesser panda is the sister group to all canoid carnivores. Thus the pandas are not a natural group. (By Dafila K. Scott.)

take blood samples from the victim, the two suspects, and an AIDS patient whom you are certain had nothing to do with the rape case. From those blood samples you isolate the RNA genome of the HIV virus and confirm that all four persons are infected. From the RNA you prepare a DNA copy that you sequence. The critical part of the four sequences is 30 bases long and is shown in Table 13.1.

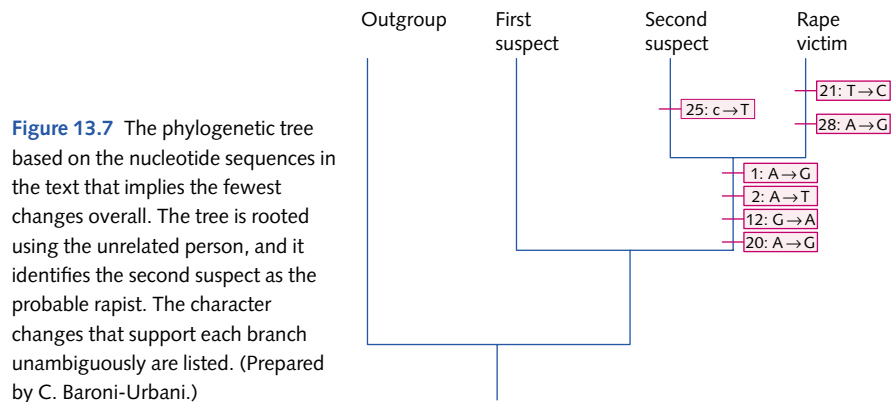
Using a computer program, you prepare a phylogenetic tree based on these four sequences. The program examines all the possible trees and delivers the one that implies fewer mutations in nucleotides than any other (Figure 13.7).

This tree suggests that the second suspect infected the rape victim. The conclusion is supported by four changes in sequence that are shared by the second suspect and the victim: A → G at position 1, A → T at position 2, G → A at position 12, and A → G at position 20. HIV is an RNA virus that evolves rapidly, and since the rape occurred, the virus has continued to evolve in both suspect and victim. In the suspect, there has been a change at position 25, C → T, and in the victim there have been two changes, at position 21, T → C, and at position 28, A → G.

This example is artificial, but similar methods were used to identify the sailor that introduced HIV to Scandinavia, the dentist with HIV who infected several

Table 13.1 DNA copies of hypothetical HIV RNA sequences taken from the rape victim, two suspects, and one unrelated person.

	1	5	10	15	20	25	30
Unrelated person	A	AGCT	TCATA	GGAGC	AACCA	TTCTA	ATAAT
Suspect 1	A	AGCT	TCACC	GGCGC	AGTTA	TCCTC	ATAAT
Suspect 2	G	TGCT	TCACC	GACGC	AGTTG	TCCTT	ATAAT
Rape victim	G	TGCT	TCACC	GACGC	AGTTG	CCCTC	ATGAT



of his patients in Florida (both the sailor and the dentist died before this was discovered), and in criminal cases to show that a gastroenterologist injected his girlfriend with blood taken from a patient with HIV (Metzker et al. 2002) and that cases of encephalitis in New York and New England were caused by the first strains of West Nile virus to appear in the western hemisphere (Lanciotti et al. 1999).

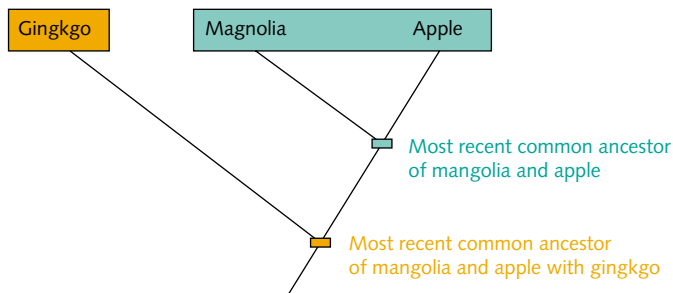
Thus systematics has discovered many surprising and important relationships ranging from the strains of HIV found in a small group of humans to the largest features of the Tree of Life. How could it do this? To answer that, we now consider the concepts, terminology, and methods of systematics.

How phylogenetic concepts are defined

Relationship is defined by how recently common ancestors were shared

For a long time after people began trying to classify organisms in a systematic fashion, they used a variety of definitions of relationship. One definition was, ‘things that look like each other are more closely related to each other than they are to things that look different.’ This is perhaps logical, but it is wrong, for some things that resemble each other superficially—such as African euphorbias and American cacti—are not closely related at all, and other things that appear to be quite different—such as the myxozoans and jellyfish discussed above—turn out to be related. Zimmermann (1931) took the critical step of defining relationship as the sharing of a recent common ancestor. For example, apples are more closely related to magnolias than they are to ginkgos because apples and magnolias share a more recent common ancestor with each other than either does with ginkgos (Figure 13.8).

By defining relationship as sharing a most recent common ancestor, we can then identify groups that are natural—groups that accurately reflect genealogy—and groups that are unnatural or incorrectly identified—groups that distort genealogy. Systematics has developed special terminology to describe natural and unnatural groups.



● KEY CONCEPT

Phylogenetic concepts are defined to be consistent with and useful in the logical discovery of natural groups, such as species and clades.

Figure 13.8 Phylogenetic relationship is defined as recency of common ancestry; the more recent the common ancestor, the more closely related the groups.

Monophyletic: All species in a monophyletic group are descended from a common ancestor that is not the ancestor of any other group and no species descended from that ancestor are not in it.

Paraphyletic: A group that does not contain all species descended from the most recent common ancestor of its members; some of those species are outside it.

Polyphyletic: A group that contains species descended from several ancestors from which members of other groups also descended.

Essential terminology includes monophyletic, paraphyletic, and polyphyletic

All species in a **monophyletic** group are descended from a common ancestor that is not the ancestor of any other group: no species descended from that ancestor are not in the group. Thus monophyly technically defines what we mean by a natural group. The dog clade, Canidae, is a monophyletic group. All of its members are more closely related to each other than they are to any species outside the Canidae, and there are no species descended from the common ancestor of the canids, to our knowledge, that are not included in the group.

A **paraphyletic** group is an unnatural group that does not contain all species descended from the most recent common ancestor of its members. One now-classical example of a paraphyletic group is the reptiles. The word reptiles refers to an unnatural group because it does not include within it the birds and the mammals, both of which are descended from branches of the phylogenetic tree located well within the reptiles (Figure 13.9). Another paraphyletic group is the fish, which does not include the tetrapods, which are descendants of fish.

Another unnatural group is called **polyphyletic** if its species are descended from several ancestors that are also the ancestors of species classified into other groups. Figure 13.10 depicts a polyphyletic group—the homeothermia or warm-blooded tetrapods—that is polyphyletic because it does not include the several groups of reptiles that share an ancestor with the birds and mammals. Other labels for unnatural polyphyletic groups are worms, algae, and protozoa.

Examples of convergence are found in plants, animals, and molecules

In addition to the terminology that describes natural—monophyletic—and unnatural—paraphyletic and polyphyletic—groups, we also need terms to describe the two major reasons that traits and DNA sequences can look the same: either because they are descended from common ancestors, or because natural selection or drift shaped them in similar ways so that they now look the same although they are descended from ancestors who were unrelated and looked different. We have abundant evidence that things descended from

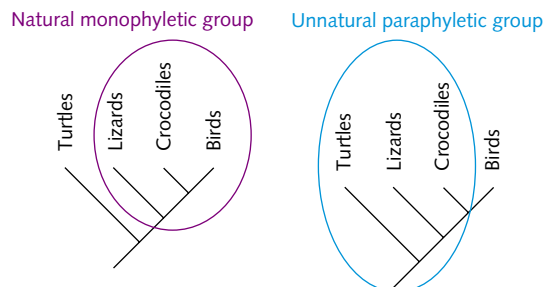


Figure 13.9 There is no common term for the natural monophyletic group on the left; the term reptiles refers to the unnatural paraphyletic group on the right, a group that does not include the birds.

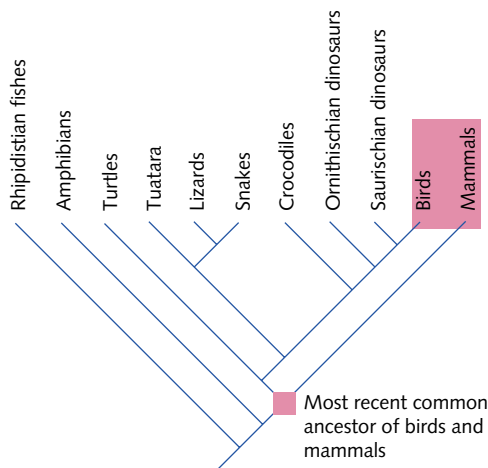


Figure 13.10 Polyphyletic groups contain members descended from several ancestors that are also the ancestors of species classified into other groups. The group homeotherms, uniting birds and mammals, is an unnatural polyphyletic group. (Reproduced with kind permission from Jonathan Armbruster, Auburn University.)

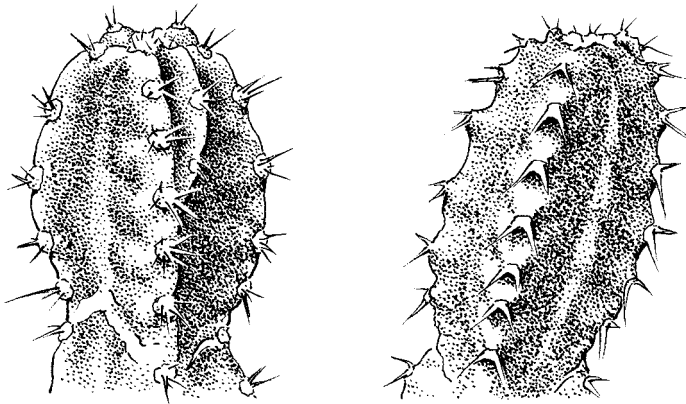


Figure 13.11 Convergence is one reason why simple resemblance is not a reliable indicator of systematic relationships. The cacti of the New World and the euphorbs of Africa have evolved similar morphologies to deal with similar ecological problems. Left: the cactus *Cereus validus* from Argentina, half size. Right: the euphorb *Euphorbia resinifera* from Morocco, double size. (By Dafila K. Scott.)

common ancestors can now look different—through evolutionary divergence—and that things descended from different ancestors can now look similar—through evolutionary **convergence**. Thus both similarities and differences can mislead us with respect to the real genealogy. Consider a few examples.

The Old and New World succulent plants, which have adapted to arid conditions with similar morphology, are a spectacular example of convergence. Some New World cacti (family Cactaceae) are so similar in shape to some Old World euphorbs (family Euphorbiaceae) that only an expert can tell them apart (Figure 13.11). In other groups flowers and fruits have converged because plants with different ancestors are pollinated by similar insects or dispersed by similar birds.

Convergence: Two species resemble each other not because they shared common ancestors but because evolution has adapted them to similar ecological conditions.

When they encounter phenotypic convergence, biologists must look at DNA sequence data or at additional morphological data to discover the underlying relationships. In animals, striking examples of convergence include the bills of Old World sunbirds, New World hummingbirds, and Hawaiian honey creepers, adapted to extract nectar from deep flowers; the fusiform shape of porpoises, tunas, sharks, and ichthyosaurs, adapted to fast swimming (see Figure 13.14 below); and the wings of birds and bats, which have independently evolved similar adaptations to rapid long-distance flight or to hovering in front of flowers.

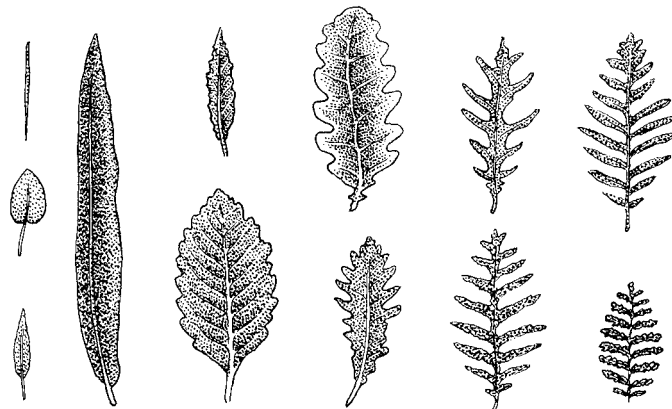
Adaptive convergence also occurs in molecules, such as the hemoglobin of birds that fly at high altitudes. The bar-headed goose lives at altitudes of 4000 m and flies up to 9200 m in the Himalayas; the Andean ‘goose,’ a relative of the mallard duck, lives at 6000 m in the Andes. Both species have hemoglobin molecules with higher oxygen affinity than those of their lowland relatives, and one mechanism of increased oxygen affinity is precisely the same in both species—disruption of a contact between the α and β chains through the same amino acid substitution in the β chain (Gillespie 1991). Thus exactly the same modification of the hemoglobin molecule occurred at least twice in distantly related species whose nearest relatives have hemoglobin molecules with different structures.

Divergence: *Related species no longer resemble each other because evolution has adapted them to different ecological conditions.*

Divergence, which occurs when originally similar species become dissimilar, is well illustrated by the Hawaiian lobelias (Givnish et al. 1995). One of the six native genera of Lobeliaceae, *Cyanea*, contains 55 species that constitute 6% of the endemic flora of Hawaii. They are restricted to particular islands or parts of islands. All *Cyanea* descend from a single ancestor; many have undergone striking changes in growth form, leaf size and shape, and flower morphology (Figure 13.12).

They vary from 1 to 14 m in height and have leaves that can be simple, compound, or doubly compound, ranging from 0.3 to 25 cm in width and up to 1 m in length. The flowers, which coevolved with endemic birds, have corolla tubes that range in length from 15 to 85 mm. The genus includes shrubs, trees, and a vine. Any classification based solely on growth form or leaf morphology would

Figure 13.12 Divergence is another important reason why simple resemblance does not indicate relationship. The Hawaiian lobelias have undergone a dramatic radiation in which their leaves have evolved many different forms. Their flowers and DNA sequences continue to indicate that they are closely related. (By Dafila K. Scott.)



not reflect phylogeny, placing members of this genus into several unrelated families. The traits that group them are flower structure, fruit color, and DNA sequence.

Among animals, the radiations of cichlid fish in the great lakes of Africa, of finches in the Galapagos and drepaniid birds in Hawaii, of land snails in Polynesia, of rodents in South America, and of amphipods in Lake Baikal all show striking recent divergence. Among plants, the huge radiation of flowering plants, and within that the radiations of the grasses, the composites, and the orchids, do the same. On a larger scale, divergence is the reason for the diversity of life.

The many cases of convergence and divergence make clear that it is not easy to see when similar traits in two species are similar because those species shared an ancestor in which that trait occurred. However, that is precisely what we need to know if we want to build reliable phylogenetic trees: we need to establish that morphological traits and DNA sequences in two or more species are similar because of shared ancestry. If that were not a problem, systematics would be simple.

Homology and orthology describe structural and molecular similarity due to ancestry

Homology is the term used by biologists to indicate that a trait in two or more species is descended from a common ancestor. Two morphological structures are called homologous by morphologists if they are built by the same developmental pathway and share the same relative position to other structures, such as nerves and blood vessels (see Figure 13.13). The hypothesis for their similarity is derivation from a common ancestor from which similar developmental mechanisms were inherited. Two genes are called orthologous if they have DNA sequences so similar that it is very likely that they derive from a common ancestor. Such molecules are similar by inspection, orthologous by hypothesis. The determination of **orthology** is more reliable for long DNA sequences, where it is very improbable that random mutation would yield similar states in two organisms. It is less reliable for short sequences.

There can be a connection between molecular orthology and morphological homology, but the connection is neither necessary nor reliable. During evolution genes can acquire new roles in new structures, and cases are known where structural homology has been preserved over long periods of time while both DNA and protein sequence homology have been destroyed (recall the *Tetrahymena* example; Chapter 7, p. xx). When DNA sequence similarity and morphological homology have different phylogenetic patterns, the difference tells us to look for something interesting in the evolution of development (Wagner 1989).

Convergent structures are analogous, not homologous

Just as homology describes underlying similarity despite divergence, **analogy** describes superficial similarity despite lack of common ancestry (Figure 13.14).

Homology: An hypothesis that similarity of a trait in two or more species indicates descent from a common ancestor.

Orthology: An hypothesis that the similarity of DNA sequences is explained by ancestry.

Analogy: Convergent traits whose similarity is caused by shared selection.

316 13 Phylogeny and systematics

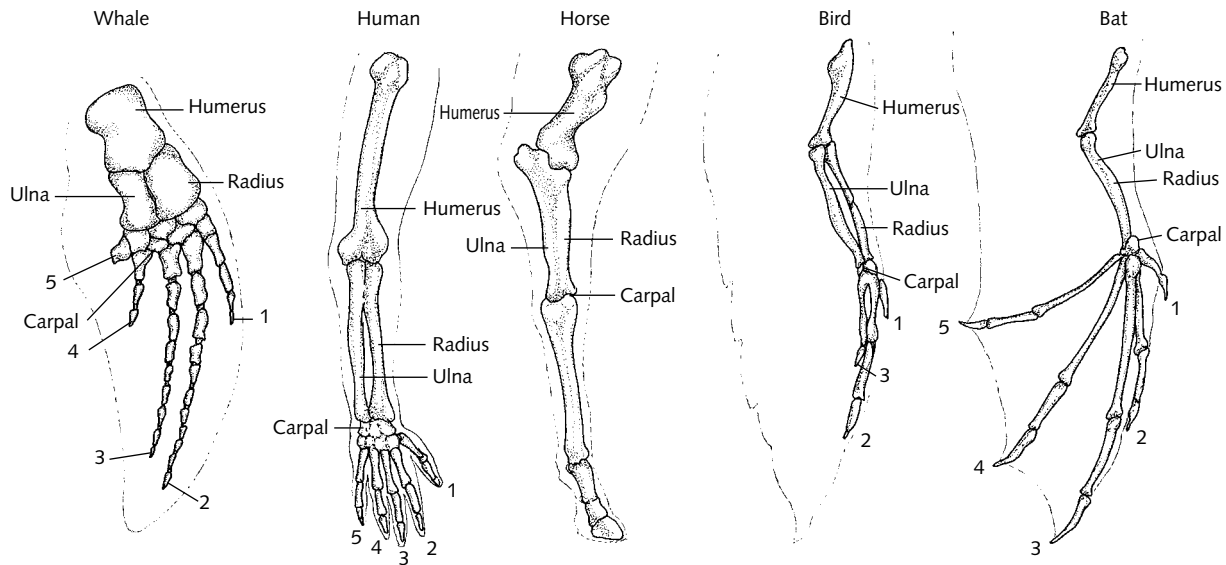


Figure 13.13 The different elements of the vertebrate limb are a classical example of morphological homology. The underlying similarities in construction can be recognized despite divergence into arms, paddles, wings, forelegs, and fins.

AQ: Please provide the source for Fig. caption 13.13, 13.14 & 13.15

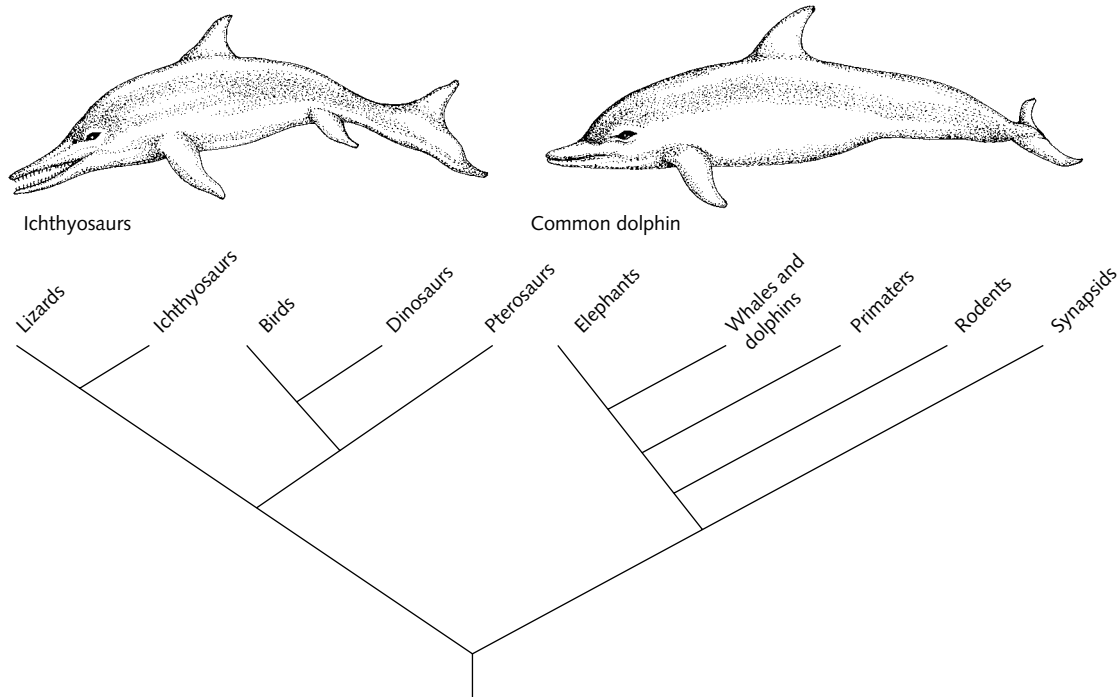


Figure 13.14 Dolphins and ichthyosaurs have similar fusiform shapes, fins and flippers, and jaws filled with teeth adapted for rapid swimming and catching fish and squid. The members of the many lineages between them do not have these adaptations. Although they look similar, they are not closely related.

It is a synonym for morphological convergence, which is one reason for **homoplasy**, a term broader than analogy that indicates similarity for *any* reason other than common ancestry, including drift. The cacti and euphorbs discussed above exemplify analogous shapes; so do the dolphins and the ichthyosaurs, which share similar shapes, live birth, and precocious offspring.

With that introduction to some important phylogenetic concepts, we now turn to tree building.

Homoplasy: *Similarity for any reason other than common ancestry. The commonest cause of homoplasy in morphological traits is convergence, in DNA sequences mutation.*

How to build a phylogenetic tree

It is clear from the examples above that phylogenetic trees can give us many insights. But how are they properly constructed?

To build a phylogenetic tree, you first need a collection of specimens of all the species in the tree. For each of these specimens, you must determine the states in which many traits are found. Those trait states are called characters.

Note that we assume that you are able to recognize the same traits in different species. This is not a problem when you are working with a collection of butterflies, but it is a problem if you are working with a collection so broad that it includes, for example, elephants and trees. At that scale, DNA sequences are much easier to interpret than morphological traits.

● KEY CONCEPT

To infer relationships properly, one must measure many characters accurately for all the species concerned, establish the homology of those characters, use them to construct all the phylogenetic trees that are consistent with them, then choose the tree (or trees) that satisfies a logical and agreed-upon criterion, of which there are several.

First construct a matrix with species in rows and characters in columns

From such information you build a character matrix with, for example, species in the rows and characters in the columns. Now you need to decide which characters tell us something useful about relationships—which characters are, in that sense, informative. One of the most important insights of systematics is that it is only shared derived characters that are informative: characters that are shared by all the species in the group you are focusing on—the **ingroup**—but not by the species in their relatives—the **outgroup**—because they originated on the branch of the phylogenetic tree leading to the ingroup. Characters that are shared between the ingroup and the outgroup because they originated in the more distant past are not informative about the relation of this group to the outgroup: there will be more about this below.

Ingroup: *An assumed monophyletic group, normally made up of the taxa of primary interest.*

You would therefore like to know which characters are derived, or relatively recent, and which characters are primitive, or relatively ancient. But you cannot make that determination of what is primitive and what is derived until you have a phylogenetic tree on which to map the characters, and the tree is precisely what you are trying to get. Thus a dilemma: without a tree, you cannot determine what came earlier and what came later; and without knowing what came earlier and what came later, you cannot describe the tree of relationships.

Outgroup: *One or more taxa assumed to be phylogenetically outside the ingroup.*

Then choose either the simplest or the most likely tree corresponding to this matrix

There are several ways out of this dilemma (Holder and Lewis 2003), as follows.

Parsimony: *The principle that things should be kept as simple as possible.*

Character state: *The specific value taken by a character in a specific taxon.*

Maximum likelihood: *A method of inferring the process that would make the data observed the most likely of all possible data sets.*

Bayesian inference: *A method that focuses on the likelihood of observing one thing given that another thing has already occurred.*

- Try all possible trees and choose those that are simplest, those that imply the fewest changes in characters. This is the principle of **parsimony**, a principle of logic called Ockham's razor and stated by William of Ockham (d. 1347): 'Entities should not be multiplied unnecessarily.' In systematics this translates into the criterion that the best tree is the one with the fewest changes in **character states** and the least convergence.
- Choose the tree that would make it most likely that you would observe the characters that you actually did observe. This is the principle of **maximum likelihood**. To use it, one must assume a model of how evolutionary change occurs; the basic model assumes the same type of random change down all branches of the tree.
- Choose the tree that would make it most likely that you would observe these branches and branch lengths, and character distributions, given a prior expectation of what the tree should look like. This approaches the problem of phylogenetics using **Bayesian inference**, a statistical method that first establishes a basic expectation (the prior probability), and then estimates the likelihood of observing the data given that expectation (the posterior probability).
- Use various combinations of these three basic ideas.

We now explore these issues in greater detail.

Synapomorphies contain information about relationships; symplesiomorphies do not

Synapomorphy and symplesiomorphy are essential systematic terms derived from Greek roots. We introduce them through the classical example of morphological homology, the vertebrate limb. Having a forelimb with humerus, radius, ulna, carpals, and metacarpals does not help to distinguish bats from turtles, because within the tetrapods that complex trait is shared among all groups, telling us nothing about their ancestor-descendent relations. However, having a limb with that complete structure does help to distinguish tetrapods from lobe-finned fishes, for in that context it is a shared, derived trait, a **synapomorphy** (shared = *syn-*, derived = *-apo-*, trait = *-morph*: synapomorphy) or a trait that originated once in their common ancestor, is shared by all of them, and is not found in their closest relatives. If some member of the group does not have it, then it is because it has been lost since the group originated. Figure 13.15 distinguishes between informative, shared,

Synapomorphy: *A shared, derived character state indicating that two species belong to the same group.*

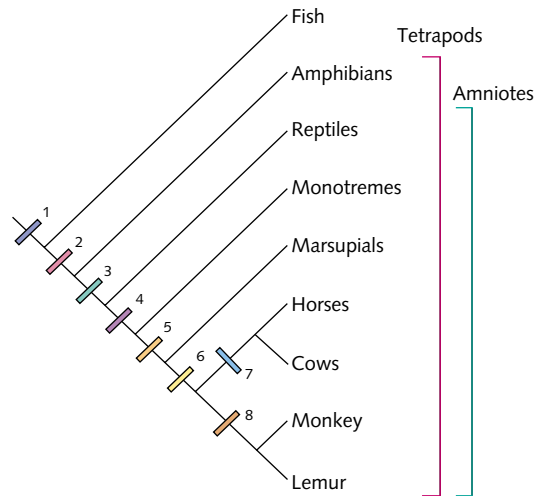


Figure 13.15 Eight key traits are informative about the evolution of a segment of the vertebrate clade: (1) the vertebral column, (2) lungs, (3) an amniotic egg, (4) lactation, (5) eggshell absent, (6) complex placenta, (7) hooves on finger and toe tips, (8) opposable thumb and fingernails. The vertebral column is a synapomorphy of the vertebrate clade but a symplesiomorphy of every vertebrate group—fish, amphibians, reptiles, and so forth. Lungs are a synapomorphy of the tetrapod clade but a symplesiomorphy of the amniotes.

derived traits—synapomorphies—and uninformative, shared, ancestral traits—**symplesiomorphies** (shared = *syn-*, recent or ancestral = *ples-*, trait = *-morph*).

There are three important points here:

- (1) simply looking similar is not necessarily informative;
- (2) informative traits are shared, derived traits;
- (3) what is shared and what is derived, and therefore what is informative, depends on the context, on what part of the tree you are looking at.

Symplesiomorphy: A character state shared by all members of this group as well as with members of related groups.

To infer the tree from the character matrix follow changes in character states between species

We start to build a phylogenetic tree by considering the elemental step—the change of a trait from one state to another (Figure 13.16). We denote the things whose relationships are being analyzed—the genes, species, or larger groups—by capital letters, A, B, C . . . , and the traits being used to determine relationship by numbers, 1, 2, 3

The problem is how to infer the correct tree from the character matrix. Figure 13.17 shows how a tree is associated with a character matrix. Each change in character state on a tree is associated with a coded change in a character matrix.

We begin not with the tree, which we are trying to infer, but with the character matrix, which we can observe. Given the character matrix depicted in

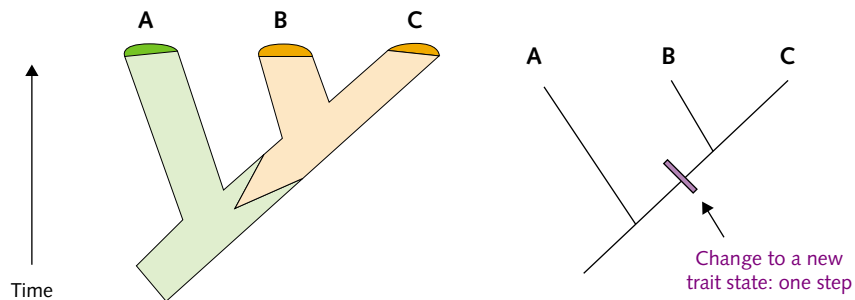


Figure 13.16 The elemental event in systematics is the evolutionary change of a single trait from one state to another. This is normally marked on the tree with a bar at the point where it happened (see Figure 13.15) and is coded 0 for the ancestral state and 1 for the derived state. Note the simplification of a complex microevolutionary dynamic (upper left) into a single step (lower right). A, B, and C are things related by ancestry—genes, species, or larger groups.

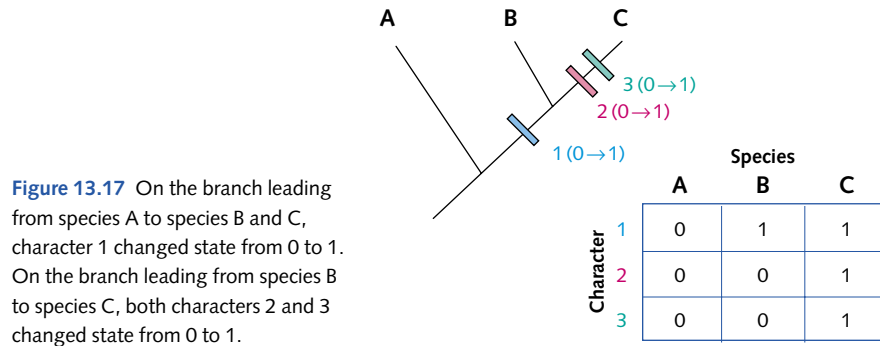


Figure 13.17 On the branch leading from species A to species B and C, character 1 changed state from 0 to 1. On the branch leading from species B to species C, both characters 2 and 3 changed state from 0 to 1.

Figure 13.17, there are two possible ways to draw the tree (Figure 13.18). The first (Figure 13.18a) emphasizes overall similarity, including the shared ancestral states (symplesiomorphies) indicated as 0s, and the second (Figure 13.18b) emphasizes derived similarity (synapomorphies) indicated as 1s.

In Figures 13.17 and 13.18 the characters changed consistently and it was possible, using a clear set of rules, to build a single best tree from the character matrices. But what should we do when the characters conflict, when they appear to be telling different stories, as is the case in the matrix in Figure 13.19, where character 3 conflicts with characters 1 and 2? To deal with character conflict we can apply the principle of parsimony, which means looking at all ways of placing the characters on all the possible trees, then choosing the tree or trees that imply the fewest changes. In this case, there is one tree that implies four changes (two different ways), and one tree that implies five changes (Figure 13.20). We choose the tree that implies fewer changes, the tree that groups B with C.

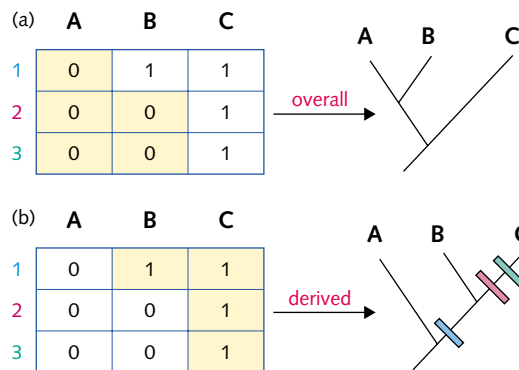


Figure 13.18 (a) If we use overall similarity as a criterion for relationship, then the ancestral states of all three traits, coded 0, predominate, and we judge A to be more similar to B than either is to C. (b) If we use derived similarity, coded 1, as a criterion, then we correctly group B with C. Note that we have indicated the emergence of the derived traits by placing markers on the tree in (b).

Trees are rooted by the choice of outgroup

After we have decided either which tree is the simplest, using the principle of parsimony, or which tree would have been most likely to produce the pattern actually observed, using the principle of maximum likelihood, there is still one more critical step to be taken. We have to choose an outgroup, a group closely related to the entire clade we are analyzing. By choosing an outgroup, we are then able to root the tree (Figure 13.21). For example, an appropriate outgroup for a phylogenetic analysis of the flowering plants would be a conifer. Rooting the tree establishes the direction of character change within the ingroup; the character states in the outgroup are assumed to be ancestral, and changes in character states within the ingroup are taken as derived by comparison to the ancestral state. Note that phylogenetic trees are like mobiles—think of them as objects that you can pick up by the root and spin freely in the air as they hang from your hand. For example, in the tree on the right in Figure 13.21, it makes no difference whether A is above B or B is above A. Both express the same relationship.

Maximum likelihood methods find the tree most likely to produce the data observed

In the maximum likelihood approach a tree is judged by how well it predicts the observed data; the best tree is the one with the highest probability—the greatest likelihood—of producing the observed pattern. To use this method, you need a way to calculate the probability of a data set given a phylogenetic tree. For sequence data, this is usually done using a model based on the probabilities of point mutations occurring at random. The most likely tree is found by considering many candidate trees—all possible trees or a representative sample of

	A	B	C
1	0	1	1
2	0	1	1
3	1	1	0

Figure 13.19 A case of character conflict, which occurs frequently. Character 3 conflicts with characters 1 and 2. Characters 1 and 2 group species B and C together; character 3 groups species A and B together. We assume that the common ancestor was in state 0 for all three traits.

Outgroup comparison: A method used to root phylogenetic trees and thus to establish the direction of character change.

AQ: Please Check term Glossary “Outgroup Comparison” not in text.

Figure 13.20 Using parsimony to decide which tree is the best when characters are in conflict. We assume here that the ancestor was in state 0 for all three traits. We choose the tree that implies the least change, in this case the tree that groups B with C, not the tree that groups A with B.

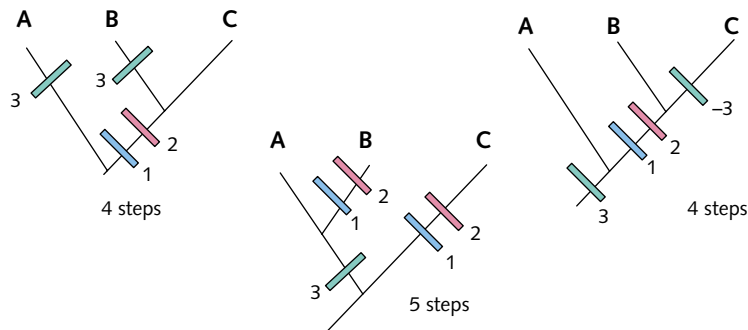
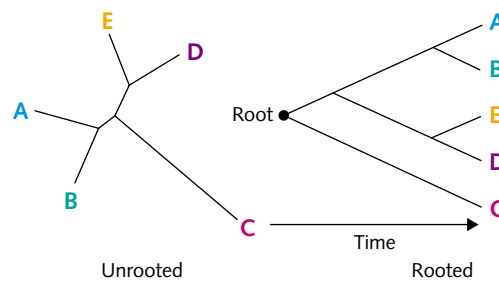


Figure 13.21 We root a phylogenetic tree by including in the analysis an outgroup, in this case C. Rooting the tree establishes polarization—what we infer as ancestral and what we infer as derived.



them. For each candidate tree the probability is calculated of finding any two sequences at opposite ends of a branch (for example a T at one end and a C at the other). This is done for all branches, the probabilities are multiplied together to get the likelihood for the whole tree, and the best tree is then the one with the maximum likelihood (Felsenstein 1988). The method is logically appealing and computationally expensive. Its range of application is increasing as computers improve.

Constructing a tree for even a modest number of species requires enormous computations

Thus far it appears that constructing a phylogenetic tree is a simple task. You get a matrix with species in columns and characters in rows, place all the characters on all the possible trees, and either pick the simplest or the most likely. It is true that the task is logically straightforward, but it rapidly becomes a computational nightmare as the numbers of species and traits increase (see Table 13.2).

To put the numbers in Table 13.2 in perspective, there are thought to be about 7–10 million species alive on the planet today, the number of protons, electrons, and neutrons in the universe is on the order of 10^{130} , and the number of trees that can be evaluated with a reasonable number of characters running a computer for 9 months in the year 2001 was about 24 billion, enough to

Table 13.2 As the number of taxa increases, the number of possible trees explodes.

Number of taxa	Number of possible binary trees
3	1
4	15
10	34 459 425
20	8 200 794 532 637 891 559 375
500	$1.0084917894 \times 10^{1280}$

analyze a tree with 10 species but not enough to handle even 20 species. The computational burden can be somewhat reduced by using a parsimony tree as the starting point for a maximum likelihood analysis, but computation remains a serious problem. If quantum computing becomes a reality, it may become possible to take the brute-force approach and evaluate all alternatives even for trees with thousands of species. Until then, we shall have to be satisfied with various approximations for trees with more than 15–20 species.

The names of groups should reflect relationships

How to construct a phylogenetic tree is one issue, and it is separate from the issue of how best to name the groups that are implied by the structure of the tree. Suppose now that we have constructed a tree and want to give names to the various groups that we can recognize on it. It is important that the names correspond to natural groups, that they reflect genealogy. Names should not be given to paraphyletic or polyphyletic groups, for they would then imply that things existed which in fact did not—they would be unnatural names. Unfortunately, the naming of names occurred before modern systematics had clarified the underlying logic of relationships, leaving us with terms like reptiles for a paraphyletic group and homeotherms for a polyphyletic group. Figure 13.22 illustrates, on the left, the traditional names for various vertebrate clades and, on the right, technical terms that properly recognize their genealogical relationships. For example, fish is not a proper phylogenetic term because it does not describe the entire clade, including tetrapods. Similarly, reptiles is not a proper phylogenetic term because it does not include all the other amniotes, including mammals.

● KEY CONCEPT

The names given to groups have not always kept up with the changes that modern systematics has made in the Tree of Life. Names should correspond to natural, monophyletic groups.

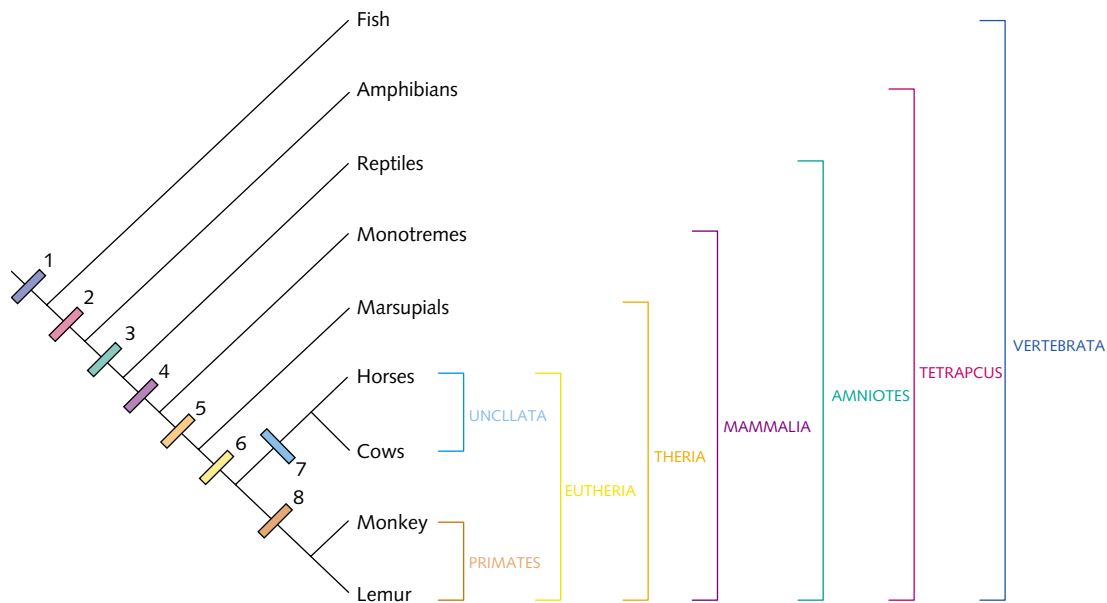


Figure 13.22 A rough outline of the vertebrate clade showing the informal and sometimes logically incorrect names of the groups on the left and the technically correct terms for the larger groupings on the right. (From Freeman 2002.)

Important issues in molecular systematics

Alignment must be used to establish the homology of sequences

KEY CONCEPT

DNA sequences are a powerful tool for identifying related species, but molecular systematics must be approached critically to be used reliably.

Because descendants inherit traits from their ancestors through genes, the history of descent is recorded in changes in the DNA sequences. Molecular data on sequences in genes are a simple form of character data: the characters are positions in the sequence, and the character states are the nucleotides at those positions. This sounds simple but assumes that the positions compared are homologous, that they derive from the same positions in a common ancestor. There are two problems with this assumption.

First, with four nucleotides, the probability that two nucleotides are the same simply because of mutation is high.

Second, in all but the most highly conserved sequences, insertions and deletions have occurred since the species being analyzed diverged from their shared ancestor. These insertions and deletions cause the overall lengths of orthologous DNA sequences in related species to differ. They also remove and add nucleotides at various places within the sequences. To make the sequence homology consistent along the entire length of the sequences being compared, gaps have to be inserted into the sequences where deletions have occurred. Gaps also have to be inserted in the sequence of one species opposite the positions where insertions have been added to the sequence of the other species. This is done so that the rest of the positions—the majority of them—that are thought to be homologous can

be aligned into the same column. Only after that alignment has been done can we compare nucleotides at the same position to see if any of them have changed.

This alignment process will be done differently depending on the assumption one makes about the ancestral sequence from which the observed sequences are thought to be derived. Thus alignment is a critical step that involves assumptions about homology and phylogeny. There are algorithms that align sequences automatically, introducing some objectivity, but the selection of an algorithm can be subjective, and the algorithms are not always reliable. In practice many alignments are performed manually.

Even after sequences have been aligned, homoplasy—sequences that are similar not because of shared ancestry but because of some process that has occurred since they last shared ancestors—remains common. Homoplasy can be reduced but not eliminated by selecting and weighting characters. Because of the problem of homoplasy in sequence data, methods based strictly on parsimony cannot extract all the information available in sequences. The desire to gain access to that additional information is one reason that molecular systematists often use maximum likelihood methods.

Thus aligning DNA sequences is equivalent, at the molecular level, to establishing the homology of morphological characters at the phenotypic level.

The neutral model and the molecular clock

The dating of lineages starts with a fossil whose age marks, at least approximately, the divergence of the lineages. Many methods in molecular systematics assume that mutations are then fixed at the same overall rates in each lineage. This assumption connects evolutionary genetics (Chapters 3, 4, and 5) to systematics. The important part of the assumption is the regular rate of substitution of nucleotides. Molecular evolution does not have to be neutral for the methods to work, but it does need to have a strong form of statistical regularity that is most plausibly supplied by neutrality. Mutations occur independently, different nucleotides are fixed in each lineage, and as time goes by, differences in sequences accumulate. That is not controversial.

What is controversial is the assumption of a **molecular clock**: the claim that each lineage accumulates changes in sequences—substitutions, or mutations that have been fixed—at the same rate. The number of changes that have accumulated then estimates the time elapsed since the lineages shared common ancestors.

There are several problems with this idea. First, the genomes of all organisms within a group share a similar structure that determines both the mutation rate and which parts of the genome are exposed to selection. Thus the rate of nucleotide change should be similar within large groups—eukaryotes, prokaryotes, RNA viruses—but not between them. The clock should tick at different rates in groups with different genomic structures.

Second, lineages differ in generation time. Small organisms usually have shorter generations than large ones, and body size varies dramatically among lineages. Since the mutation rate is a rate per generation, not a rate per year, one

AQ: Please check definition of “molecular clock” not available in the typescript.

would expect changes to accumulate more rapidly in lineages with short than with long generations. This effect has been demonstrated in several cospeciating groups, such as gophers, with longer generation times, and their lice, which have shorter generation times.

Third, if the group being analyzed met all the assumptions required for a molecular clock, then the phylogenetic tree produced by the analysis would be clock-like, meaning that the distance from the common ancestor to the tips of each of the branches would be the same: each path would have the same number of codon or amino acid substitutions. However, this is rarely the case. The lengths of the branches connecting one ancestor to several descendants often differ. One might want to accept a rough correlation between divergence times and number of differences in substitutions between lineages, rather than a precise fit to a clock-like tree. However, even there problems arise, for the confidence limits that one can place on such relations are so broad that in practice the resulting clocks are imprecise.

For all these reasons, many studies try to avoid the assumption of a molecular clock. However, they are used when no other method is available to estimate the timing of an event, and when compared with the times delivered by fossils, the times delivered by molecular clocks often suggest interesting disparities. Recall that orthologous sequences are sequences in two or more related species derived from what was a single sequence in the most recent shared ancestor. A good predictor of the amount by which orthologous sequences in two species have diverged remains the time that has elapsed since they existed in a common ancestor. A longer branch in a molecular phylogeny suggests that more time has elapsed along that branch than along a shorter one—except, of course, when they are sister-branches.

Use the right molecule for the problem at hand

Each type of molecule and method of analysis is best suited to a certain range of problems. Hillis et al. (1996) and Avise (1994) discuss in depth the advantages and disadvantages of various molecules and methods. Here we simply make the point that one should choose the right combination for the problem at hand.

Molecules are like radioisotopes: they change at different rates. Uranium²³⁸ has a half-life of 4.5×10^9 years, which makes it useful for dating objects of about the age of the Earth or the moon, whereas carbon¹⁴ has a half-life of 5730 years, which makes it useful for dating archaeological objects from a few hundred to about 20 000 years old. The genomes of RNA viruses like HIV change so quickly that every person infected soon carries an identifiably different strain. Mitochondrial DNA, which is haploid, has a relatively fast substitution rate. It evolves rapidly enough to be useful for comparisons of lineages that diverged recently, but it can also be used to establish relationships among groups that are several million years old. Beyond that point it becomes so altered by repeated mutations that useful information is obscured by noise.

To get good molecular information on events that occurred in deep time, we need highly conserved genes, genes that change very slowly, such as the DNA

that codes for the small subunits of ribosomal RNA. Such genes contain useful information about events that occurred 500–1500 million years ago. They can be used, for example, to test the idea (Margulis 1970, 1981) that mitochondria and chloroplasts are intracellular symbionts derived from prokaryote ancestors, an idea now strongly supported by sequence data:

- the nucleotide sequences of the 16 S RNA gene from chloroplasts indicate that chloroplasts are more closely related to photosynthetic cyanobacteria than to the nuclear genome of maize;
- similar analysis suggests that mitochondria are derived from the α subdivision of the purple bacteria.

Comparing the phylogenies of organelles and nuclei reveals striking differences in their geometry (Figure 17.7). The nuclear sequences support the traditional view that the plants, fungi, and animals form a group distinct from the protists. The mitochondrial sequences suggest that the plant mitochondria were independently derived from the purple bacteria much more recently than the mitochondria found in fungi, ciliates, green algae, and animals. Chloroplasts appear to have been more recently acquired than mitochondria, for their sequences have diverged less from their bacterial ancestors, less of the chloroplast than of the mitochondrial genome has been transferred to the nuclear genome, and they entered hosts that did not yet exist when mitochondria entered the eukaryote lineage.

Thus highly conserved DNA sequences record the ancient history of organisms that have left no fossils, from a time in Earth history when we have very little other information, and allow us to test a major evolutionary hypothesis, the symbiotic theory for the origin of eukaryotic cells.

The genealogy of genes can differ from the phylogeny of species

Genes in the same organism can have different evolutionary histories

Molecular systematics is not only used to build trees relating species; it is also used to construct the history of single genes. The trees constructed from different genes in the same organisms often have different structures because each gene has had a different evolutionary history. A gene genealogy can differ from a species phylogeny because mutations do not occur simultaneously and are not constrained to occur during speciation (Figure 13.23). One gene may have diverged prior to a speciation event, another gene may have diverged after that speciation event. Thus genes have different genealogies, and only some genealogies have the same structure as the phylogeny of the species in which the genes occur. For events occurring within a species, the recovery of a reliable gene genealogy must be done in sequences with little or no recombination, such as the mitochondrial genome, because recombination produces nets, not branches.

● KEY CONCEPT

Because mutation is not synchronized with speciation, genes and species have histories that are similar in general but can differ in important details.

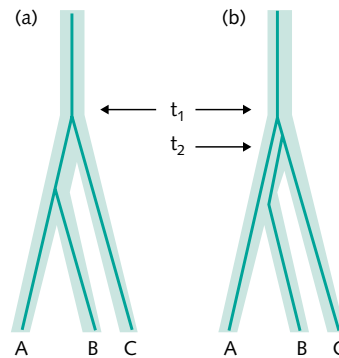


Figure 13.23 Phylogenies of species and genealogies of genes: The species tree is described by the large outer figure and is the same in both cases. The gene genealogy is described by the lines contained within the tree. Whenever more than one line is present, the gene is polymorphic. In both lineages there is a first mutation prior to the first speciation event. (a) The gene tree and the species tree have the same branching pattern, i.e. the second mutation occurs between the first and the second speciation events. (b) The genes and species have different branching patterns. The second mutation event occurs shortly after the first mutation event and before the first speciation event; thus here the genealogical split pre-dates species divergences. (From Avise 1994.)

A time to the most recent common ancestor is estimated from the difference in DNA sequences

Consider the DNA sequences of two copies of the same gene. They could be two alleles from a single population or from two related species. Assume that there has been no recombination, and that mutations have been neutral. That would be the case if the DNA were sampled from mitochondria or other haploid asexuals and if mutations produced no change in protein function. Such changes should be neutral or nearly so.

The two copies of the gene differ at several neutral sites. At some time in the past, when both copies of the gene derived from a common ancestor, there were no differences between them. How long did it take for this many differences to accumulate? To make that calculation, we must assume a constant mutation rate, but we do not have to make any assumptions about population size or selection at nearby loci, for neutral mutations accumulate within genes at rates that do not depend on these factors. Other important genetic properties of the population do depend on population size and selection; these include the number of mutations that will be fixed in the entire population and the amount of polymorphism that exists in the population at any time. But the number of mutations that have been fixed along an individual lineage since the last common ancestor depends only on the mutation rate and the time elapsed (Hudson 1990).

Mutation rates for single nucleotides are about 10^{-8} to 10^{-9} per organism per generation. With that information, we can use the neutral theory of evolution (Chapter 3, p. xx) to estimate how long ago the common ancestor existed. The error in the estimate depends on the lengths of the DNA sequences and the

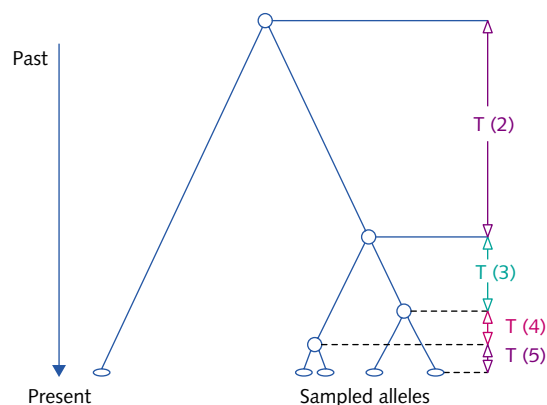


Figure 13.24 In coalescence analysis one uses a mathematical model of evolution that assumes a similar rate of fixation of mutations in all branches of the tree, and calculates backwards from an array of existing species until all branches of the tree coalesce in a single root. $T(5)$ is the estimated age of the youngest allelic relatives, $T(4)$ the next youngest, and so forth. (From Hudson 1990.)

number of mutations detected in them. We build the tree starting from the tips of the branches in the present, then work back in time, calculating the points of **coalescence** of the branches into a common ancestor. The trees that result are not phylogenies of species but genealogies of genes, and the process is called a coalescent process because the calculations yield the age at which the differences coalesce into the same ancestral sequence (Figure 13.24).

Once a reliable genealogy is obtained, the number of mutations separating the molecular ancestor from its tips, and the assumption of a molecular clock, can be used to date when the most recent common ancestor lived. Reliable genealogies allow us to compare old and young polymorphisms. The age of polymorphisms can be either greater or less than the age of the common ancestor of two species.

We have here described only the simplest possibility—no recombination, no geographical structure, and no selection—but the methods have been extended and can, to some extent, deal with more-complex scenarios.

Coalescence: A process of inference from different existing DNA sequences in related species back to the single ancestral sequence of the shared ancestor.

● SUMMARY

This chapter discussed what phylogenetic trees are, how to construct them, and some issues that arise in their construction.

- The Tree of Life is not obvious, it must be discovered. Because much of the information that we need to discover it has been modified by the evolutionary process, we have to use methods that compensate for those modifications.
- Modern phylogenetic methods are making many changes in traditional views of the Tree of Life, some at the largest scale, such as the discovery of Archaea as the third major group of life, and some at the smallest scale, such as changes in which species is considered to be the closest relative of some other species.

- Reliable phylogenetic methods are based on using characters that are informative about relationships. The informative characters are shared, derived characters, not characters that are shared because they were present in distant ancestors and have not changed since.
- The starting point for a phylogenetic analysis is a character matrix—a matrix or table that lists the series of species to be analyzed and the characters being used. Many trees can be built from the same character matrix. We prefer either the simplest, the one implying the least change, or the tree that maximizes the probability of observing the distribution of character states actually seen, or some combination of these and other criteria.
- Trees are built from data using methods that can produce different trees. If all methods yield the same branch of a tree in a large and reliable data set, then that branch can be regarded with confidence. If the data are equally consistent with several branching patterns, judgement about relationships should be suspended.
- Genetic drift in microevolution produces a rough molecular clock in macroevolution. If a tree is clock-like, then the length of a branch is proportional to the time that elapsed along that branch. This is often roughly the case but rarely precisely the case.
- The genealogy of genes also produces trees that trace the history of genes, but those trees may differ from the phylogenetic trees of species within which they are embedded.

Given a reliable phylogenetic tree, we can then conduct a comparative analysis of trait evolution and historical biogeography, the topics of the next chapter.

● RECOMMENDED READING

Hillis, D.M., Moritz C., and Mable, B.K. (eds) (1996). *Molecular systematics*, 2nd edn. Sinauer Associates, Sunderland, MA.

Kitching, I.L., Humphries, C.J., Williams, D.M., and Forey, P.L. (1998). *Cladistics: the theory and practice of parsimony analysis*. Oxford University Press, Oxford.

● QUESTIONS

- 13.1. Why is it more reliable to interpret fossils as the tips of dead branches than as the direct ancestors of living species?
- 13.2. Can a phylogenetic tree ever be anything more than just a working hypothesis?
- 13.3. If every branch in the Tree of Life defines a new natural group, and if the Tree of Life includes everything from bacteria to whales, then how useful is the Linnean system of taxonomy, which tries to place all organisms into nested categories, species within genera within families within orders within classes within phyla? If that system does not work well, what would you recommend to replace it? A taxonomy should accurately reflect all the information in a phylogenetic tree. Is that goal possible?