2 Tools of the Trade

In this book we will focus on the use of the comparative method to unveil patterns of biological diversity in ecological and ethological settings and to help study the mechanisms underlying these patterns. Although comparison is one of the cornerstones of biological investigation, the "comparative method" has been approached from several directions (e.g., Ridley 1983; Clutton-Brock and Harvey 1984; Bell 1989). For one group of researchers this method consists of comparing distantly related species living in a common environment and explaining any similarities as convergent adaptations to that environment. For other researchers, the comparative method embodies reconstructing the phylogenetic relationships within a group of species, then using those relationships as a template for explanations about the evolutionary origin and diversification of other characters (see also Coddington 1988). The attraction of the first approach lies in the assumption that the evolutionary history of the test groups is unimportant. However, because the decision that two groups are "distantly related" is often highly subjective, this approach may fail to distinguish similarity due to common environments from similarity due to common inheritance. The second approach will distinguish various kinds of similarity but requires a direct estimate of phylogeny. Since a robust methodology for estimating phylogeny is available in the postulates of phylogenetic systematics (Hennig 1950, 1966; see also Eldredge and Cracraft 1980; Nelson and Platnick 1981; Wiley 1981, 1986a,b,c,d,e,f), we advocate using the second form of the comparative method because of its greater explanatory scope.

Comparative biologists often use "similarity" as an indicator of "relationship." Hennig, noting that there are different kinds of similarities and relationships, emphasized that the primary goal of systematics should be the delineation of a special type of similarity (homology) that, when used to reconstruct relationships, would provide a general reference system for comparative biology. Hennig reasoned that this system should be based on reconstructing phylogenetic relationships from shared homologous traits because **all homologies covary with each other and with phylogeny.** Other types of

similarity (homoplasy), although evolutionarily interesting, are not phylogenetically informative because **homoplasies need not covary with anything.** Evolution, with its underlying assumption of the preeminence of genealogical ties, has been the unifying biological principle since the emergence of Darwinian ideas, so Hennig's perspective would seem conceptually unobjectionable. However, response to this new perspective was reserved, originating, in part, from a long-standing problem with the relationship between homology and phylogeny. Specifically, if homology is both defined by phylogeny and required to reconstruct phylogeny, then a researcher needs to adopt an Orwellian doublethink strategy in order to "know the phylogeny, obtain the homologies, and build the phylogeny." This relationship makes phylogenetic reconstruction irreducibly circular.

Hennig's solution to this problem stemmed from his belief that genealogical influences in evolution are so pronounced that homologous characters will outnumber covarying homoplasious characters within any given group. He suggested that researchers begin with the assumption that all characters that conform to *nonphylogenetic criteria for homology,* such as those proposed by Remane (1956), are, in fact, evolutionarily homologous. In some cases this will lead to the incorrect, and initially undetectable, identification of homoplasious traits as homologues. When a phylogeny is reconstructed by grouping taxa according to their shared homologies, these misidentifications will be revealed because the homoplasious characters will not covary with the majority of the other characters. These traits can then be recognized, using the phylogenetic hypothesis, as homoplasies. The distinction between the non-Hennigian and Hennigian approaches is a subtle but vital one. Consider the following example. Suppose, while describing the behavior of four different avian taxa, you notice that all of the males perform the same type of mock-preen display during courtship. A non-Hennigian systematist would say, "Since this display looks the same in these four taxa *and* is performed by different members of a closely related group ('birds'), it is a homologous trait. We can use this trait to assess the phylogenetic relationships among these birds." A Hennigian systematist would say, "Since this display looks the same in these four taxa, it is the same (is homologous). We can use this trait to assess the phylogenetic relationships among these organisms." In the first case, homology is assumed because of similarity among characters, *coupled with* presumed relatedness among the taxa bearing the characters. Hence, there is an underlying assumption of prior knowledge of phylogeny. In the second case, homology is assumed solely on the basis of similarity among characters (the Wiley criterion "if it looks like a duck and quacks, it's a duck"). Hence, the approach advocated by Hennig is not circular because **homologies, which indicate phylogenetic relationships, are determined**

***without* a priori reference to a phylogeny, while homoplasies, which are inconsistent with phylogeny, are determined as such *by* reference to the phylogeny.**

Hennig recognized that there are three types of homologous characters; shared general characters, which identify a collection of taxa as a group; shared special characters, which indicate relationships among taxa within the group; and unique characters, which identify particular taxa within the group. In addition to these three categories of homology there is a separate category, homoplasy or "false homology," which tells us nothing about relationships among taxa. Since only shared special homologies denote particular phylogenetic relationships within a study group, characters must be assigned to one of the three homology categories before their usefulness in a phylogenetic study can be determined. Once again, the risk of circularity is high: "if two taxa are related, then a character that they share in common is a shared special homology; therefore, this character can be used to determine if the taxa are related." Hennig suggested that this determination be made by comparing the state of each character in the study group to the state of the same characters in one or more species outside the study group (outgroups). In this way, **each character is independently assigned a particular homology status (general, special, or unique) depending upon properties of species for which the phylogenetic relationships are not being assessed (the outgroups).** This "outgroup comparison" distinguishes among traits that are shared between the outgroups and at least some members of the study group (shared general traits), traits that are restricted to some members of the study group (shared special traits), and traits unique to single members of the study group (unique traits).

In the remainder of this chapter we will present a detailed discussion of the basic methods involved in phylogenetic systematics adapted largely from Wiley et al. (1991). More-advanced applications of phylogenetic systematic methodology pertinent to historical ecology will be introduced in chapters 5 and 7.

Terminology

There is a perception that researchers are required to learn an inordinate number of new and specialized words before their initiation into phylogenetics can be completed. This apprehension is based, in part, on the incorporation of numerous old terms such as monophyly, ancestor, homology, and homoplasy into the field of phylogenetic systematics. Since most evolutionary biologists are familiar with these terms, this is not really so daunting after all. There are of course some new words to learn, apomorphy and plesiomorphy being the most important and, for many researchers, the most

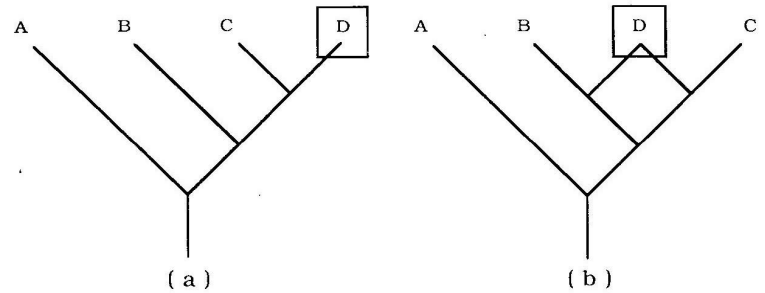


Fig. 2.1. The endpoints of speciation. *Letters* = species. (*a*) Cladogenesis: species D is produced by the division of an ancestral species into two new daughter species C and D. (*b*) Reticulate speciation: species B and C hybridize, forming species D.

perplexing. We hope that, by the end of this chapter and definitely by the end of this book, these words will no longer be fraught with such mystical connotations. There are two reasons for adopting the terminology, both the old and the new, that we will use in this book. First, the terms used in any empirical field should be as unambiguous as possible because hypotheses and explanations are framed within this terminology. Second, because evolution is a genealogical process, it is critical that systematic data be incorporated explicitly into evolutionary explanations.

Groups of Organisms

A **taxon** is a group of organisms that is given a name. The relative position (or rank) of a taxon in the Linnaean hierarchical system of classification is indicated by the use of categories (e.g., "family," "genus"). *You should not confuse the rank of a taxon with its reality as a group.* For example, the taxon Aves includes exactly the same organisms whether it is ranked as a class, an order, or a family. A **natural taxon** is a group of organisms that exists as a result of evolutionary processes. There are two kinds of natural taxa: species and monophyletic groups. A **species** is a lineage, a collection of organisms that share a unique evolutionary history and are held together by the cohesive forces of reproduction and development. Every species originates from a single **ancestral taxon** through either **cladogenesis,** the division of one ancestral species into new daughter species (fig. 2.1a), or **reticulate speciation,** the formation of a new species through the hybridization of two ancestral species (fig. 2.1b).

A **monophyletic group,** or **clade,** is a group of taxa encompassing an ancestral species and all of its descendants (fig. 2.2). Members of a monophyletic group are bound together by common ancestry relationships that they

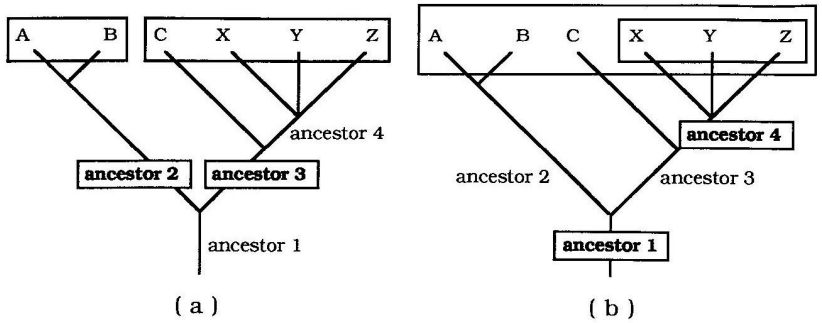


Fig. 2.2. Monophyletic groups on a phylogenetic tree. *Letters* = species. Species within boxes are part of the following monophyletic groups: (*a*) ancestor 2 and all its descendants (ancestor 2 + species A + B); ancestor 3 and all its descendants (ancestor 3 + species C + ancestor 4 + species X + Y + Z); (*b*) ancestor 1 and all its descendants (ancestors 1 + 2 + 3 + 4 + species A + B + C + X + Y + Z); ancestor 4 and all its descendants (ancestor 4 + species X + Y + Z).

do not share with any other taxa. Each monophyletic group begins as a single species, the ancestor of all subsequent members of the clade. Because of the nature of speciation, groups of species cannot give rise to other groups or to a single species. In brief, the various processes involved in speciation allow one species to give rise to another (or two species to produce a new taxon of hybrid origin), but there are no documented processes that can produce a genus from a genus ("geniation") or a family from a family ("familiation"). This occurs because species are the largest units of taxic evolution; they are real, evolutionary entities (we will discuss this in more detail in chapter 3). Higher-level categories, on the other hand, are artifacts of our propensity to classify our surroundings. As figments of our collective imaginations, supraspecific taxa have no evolutionary substance, whereas species, and the array of speciation processes that form them, lie at the very heart of "descent with modification" or evolution.

An **artificial taxon** represents an incomplete or invalid evolutionary unit. **Paraphyletic groups** are artificial because one or more descendants of an ancestor are excluded from the group, making such groupings incomplete units (fig. 2.3). For example, most researchers think that *Homo sapiens* shares a common ancestor with the African great apes (chimpanzees and gorillas). A group within a classification comprising the African great apes plus orangutans and gibbons (the Asian great apes) while excluding humans would be paraphyletic.

Polyphyletic groups are artificial because taxa that are separated from each other by more than two ancestors are placed together without the inclusion of all the descendants of that common ancestor (fig. 2.4). Since the

relationship between the two taxa is so distant, this type of grouping misrepresents the evolutionary relationships that have arisen from speciation events following the divergence of the shared common ancestor, making the grouping an invalid evolutionary unit. A classic example of a polyphyletic group would be a classification that placed bats and birds in the same taxon.

The **ingroup** is any group of theoretically closely related organisms of interest to an investigator (see fig. 2.5). *Choice of the ingroup is constrained only by the rule that it must contain more than two species* because it is

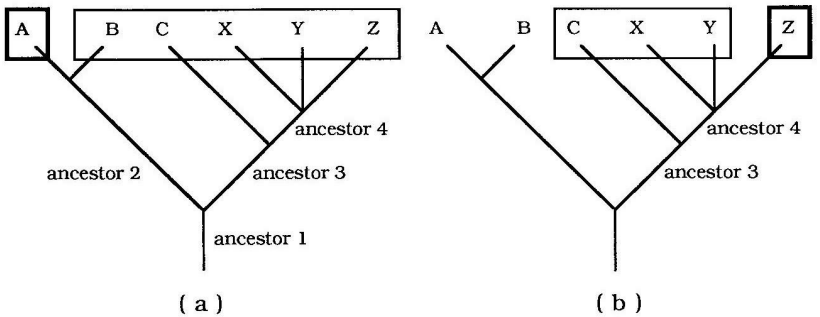


Fig. 2.3. Paraphyletic groups on a phylogenetic tree. *Letters* = species. (*a*) Two groups have been distinguished: group 1 includes ancestors 1 + 2 + 3 + 4 + species B + C + X + Y + Z; group 2 contains species A. Species A should be included in group 1 because it shares ancestors 1 and 2 with that group. (*b*) Two groups have been distinguished: Group 1 includes ancestors 3 + 4 + species C + X + Y and group 2 contains species Z. Once again, Z should be included in group 1 because it shares both ancestors with that group.

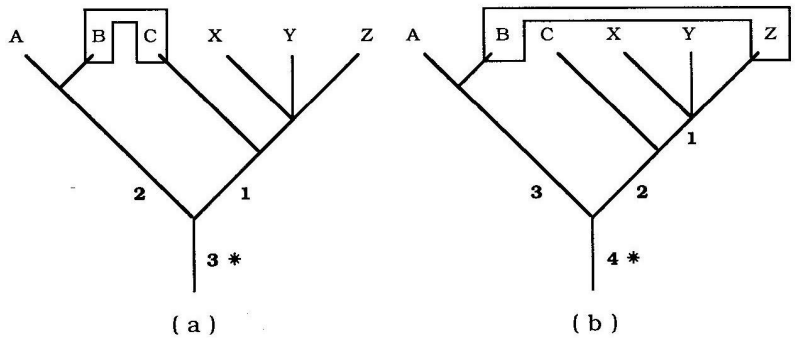


Fig. 2.4. Polyphyletic groups on a phylogenetic tree. *Letters* = species. (*a*) Species B and C are grouped together because they "look the same" even though they do not share a recent common ancestor (you have to count back through two ancestors before arriving at an ancestor, marked with an asterisk, that the taxa share). (*b*) Species B and Z are placed together; you have to count back through three ancestors before arriving at a common link between the two taxa.

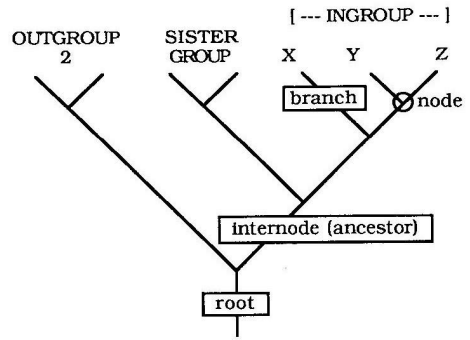


Fig. 2.5. Components of a phylogenetic tree. On this diagram, the internode is the common ancestor of the sister group and the ingroup, the node represents the speciation event that gave rise to species Y + Z, and the branch refers to species X. In all, there are six nodes, five internodes, and seven branches on this particular tree.

impossible to determine phylogenetic relationships for only two taxa. For example, an investigator studying the genealogical relationships between tigers and gorillas can only say that, by virtue of their status as biological entities, they are related. However, add lions to the picture, and we increase our degrees of freedom from none to three, because there are four possible hypotheses of relationships: they are all equally related to one another, gorillas are more closely related to tigers, gorillas are more closely related to lions, or lions are more closely related to tigers. A **sister group** is the taxon that is most closely related genealogically to the ingroup (see fig. 2.5). For example, Old World monkeys are the sister group of the great apes. The ancestor of the ingroup cannot be its sister because it is a member of the ingroup. An **outgroup** is any group used for comparative purposes in a phylogenetic analysis (see fig. 2.5). Since outgroups are used to assess the evolutionary sequence of appearance of homologous characters independently, *choice of the outgroup is constrained by the rule that it cannot contain any members that are part of the study group*. Because genealogy is so important in evolution, it is not surprising that the most important outgroup in any study is the sister group to the taxa being investigated. So, if we were interested in studying the phylogenetic relationships within the great apes, we would use the Old World monkeys as our outgroup.

Relationships of Taxa

In phylogenetic systematics the term **relationship** refers strictly to connections based on genealogy. In other systems "relationship" may be equated with "similarity" without evolutionary implications or with the implication

that taxa that are more similar to each other are more closely related evolutionarily. The latter system, based on the a priori assumption that things that "look the same are the same," can lead to the recognition of polyphyletic groups because it lacks a rigorous methodology to test the validity of the assumption. *Degree of similarity is never equated with degree of relatedness in the phylogenetic system.* **Genealogical descent** at the taxic, rather than the individual, level is based on the proposition that ancestral species give rise to daughter species through speciation. A **phylogenetic tree** is a branching diagram depicting the sequence of speciation events within a group. As a graphical representation of genealogy, *a phylogenetic tree is a hypothesis of the genealogical relationships among taxa.* Since phylogenetic trees are hypotheses and not "facts," they are dependent upon both the quality and quantity of data that support them. A tree is composed of several parts (fig. 2.5): a **branch point,** or **node,** sometimes highlighted with a circle, representing an individual speciation event; a **branch,** the line connecting a branch point to a terminal taxon, representing the terminal taxon; and an **internode,** the line connecting two speciation events, representing an ancestral species. The internode at the bottom of the tree is given the special term **root.**

Classifications

A **natural classification** contains only monophyletic groups and is thus consistent with the phylogenetic (evolutionary) relationships of the organisms. In other words, the genealogical relationships depicted on the phylogenetic tree can be reconstructed from the classification scheme. An **artificial classification** contains one or more paraphyletic or polyphyletic groups, rendering it inconsistent with the phylogeny of the organisms. In such cases the phylogenetic tree cannot be wholly reconstructed from the classification scheme. An **arrangement** is a classification of a group whose phylogenetic relationships have not yet been delineated, so it can be either a natural or an artificial classification. The overwhelming majority of current classifications are arrangements, serving as necessary but interim vehicles for classifying organisms until their phylogenetic relationships have been determined. Neither artificial classifications nor arrangements have been constructed via a rigorous, phylogenetic methodology. *It is therefore inappropriate to convert such classification schemes into phylogenetic trees, because you cannot assume a priori that taxonomic relationships are consistent with phylogenetic relationships.*

The **category** of a taxon indicates its relative place (or rank) in the hierarchy of the classification. The Linnaean hierarchy is the most common taxonomic classification scheme. Within this scheme, the formation of category names occupying specific places in the hierarchy is governed by rules con-

tained in various codes of nomenclature. It is important to remember that the rank of a taxon does not affect its status in the phylogenetic system. *All monophyletic taxa are equally important and all paraphyletic and polyphyletic taxa are equally misleading to the phylogeneticist.*

Features of Organisms

A **character** is any observable part, or attribute, of an organism. Characters have two evolutionary options: they can either remain the same and be passed on genetically from ancestor to descendant unaltered, or they can change in one species and be transmitted in the new form to its descendants. If a trait changes, it is transformed from its existing (ancestral) condition into an **evolutionary novelty.** Two characters found in different taxa can thus be assigned homologous status because they are either the same character that is found in the common ancestor or they are different characters that are genealogically linked by passing through the transformation from an ancestral condition to a novel condition. The ancestral character is termed the **plesiomorphic character** (*plesio* = close to the stem; *morpho* = shape), while the descendant character is termed the **apomorphic character** (*apo* = away from the stem; *morpho* = shape). A **homoplasy** is a character shared among taxa that is similar but does not meet either of the two preceding criteria of homology.

A **transformation series** is a collection of homologous characters: two homologous characters produce a binary transformation series, while three or more homologous characters create a multicharacter or multistate transformation series. An **ordered transformation series** is a hypothesis of the particular pathway a character travelled during its evolutionary modification(s); however, without further information we cannot tell which direction the character moved along the pathway. If a transformation series is **unordered,** there are several possible routes open to explain the character changes. Information about the evolutionary direction of character change is provided by polarization. For **polarized transformation series,** the relative apomorphic and plesiomorphic status of characters has been determined, so we have a hypothesis of which character state represents the ancestral condition and which the derived condition. And finally in **unpolarized transformation series,** the direction of character evolution remains unspecified. There are thus four possible types of character transformation series based on the amount of information available concerning the pathway and direction of evolutionary change: ordered, unpolarized (fig. 2.6a); unordered, unpolarized (fig. 2.6b); ordered, polarized (fig. 2.6c); and unordered, polarized (fig. 2.6d). Not surprisingly, ordering and polarization of multicharacter transformation series

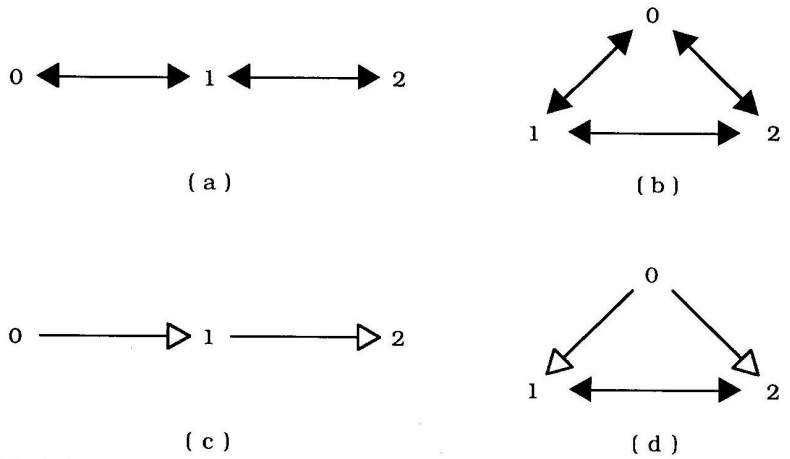


Fig. 2.6. Types of character transformation series. (*a*) Ordered, unpolarized: information about pathway; no information about direction. Character modification may be either 0 to 1 to 2, or 2 to 1 to 0. (*b*) Unordered, unpolarized: no information about either direction or pathway. (*c*) Ordered, polarized: information about both pathway and direction. (*d*) Unordered, polarized: information about direction; no information about pathway. Zero is the plesiomorphic state, but we do not know whether character modification moved from 0 to 1 to 2, or 0 to 2 to 1, or whether 1 and 2 arose independently from 0.

can become very complicated. Binary characters are much simpler because they are all automatically ordered (but not necessarily polarized!).

Character argumentation is the logical process of determining which characters in a transformation series are plesiomorphic and which are apomorphic based on a priori deductive arguments using outgroup comparison. Frequently termed "polarizing the characters," this is the pivotal process in phylogenetic systematics. **Polarity** refers to the plesiomorphic or apomorphic status of each character. **Character optimization** consists of a posteriori arguments concerning how particular characters should be polarized given a particular tree topology.

Phylogenetic systematists are quickly converting to computer-assisted analysis of their data. Such analyses require the production of a **data matrix** composed of transformation series and taxa. Each character in the matrix is assigned a numerical **code.** By convention, the code 0 is usually assigned to the plesiomorphic character, while 1 is reserved for the apomorphic character of a transformation series if the polarity of that series has been determined (hypothesized) by outgroup comparisons. If a transformation series consists of more than two characters, the situation becomes more complex, as we will discuss later in this chapter.

Hennig Argumentation: Building Trees

Phylogenetic systematists assume that all organisms, both living and extinct, occupy a unique position on one phylogenetic tree rooted at the origin of life on this planet. Since characters are features of organisms, they should have a place on this tree corresponding to the point at which they arose during evolutionary history. So ultimately we are seeking to reconstruct phylogenetic trees in which the taxa are placed in correct genealogical order and the characters are placed where they arose. For example, consider the tree for some major land-plant groups (fig. 2.7). This diagram provides us with a hypothesis of the genealogical ties between plant groups (i.e., tracheophytes and mosses are more closely related to each other than either is to hornworts). It also provides us with a hypothesis of the evolution of specific characters. Notice that characters are depicted on the phylogenetic tree at their hypothesized point of origin. This is the shorthand notation for stating that "the ancestor in which character x arose, and all of its descendants, display character x, unless it is modified again later in the evolutionary history of the group." For example, the phylogenetic hypothesis states that xylem and phloem originated in the common ancestor of mosses and tracheophytes. In other words, these characters arose in an ancestral species between the time of origin of the hornworts and the speciation event that produced the mosses and tracheophytes. Since xylem and phloem are postulated to be homologous in all plants bearing these tissues, each trait appears only once, at the level on the tree where it is thought to have arisen as an evolutionary novelty.

Now, even without a phylogenetic tree we might suspect that all plants bearing xylem and phloem shared a common, unique ancestor because both characters are morphologically and developmentally similar in these plants.

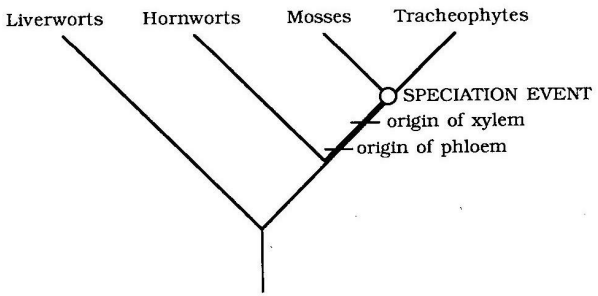


Fig. 2.7. The phylogenetic relationships of four groups of plants. Xylem and phloem originated in the ancestor of the mosses and tracheophytes. (Redrawn from Bremer 1985; cited in Wiley et al. 1990.)

In the phylogenetic system such detailed similarity is always considered as a priori evidence that the characters are homologous. This concept is so important that it has been termed

> **Hennig's Auxiliary Principle:** *Never assume convergent or parallel evolution; always assume homology in the absence of contrary evidence.*

The principle is a powerful one. Without it we could assert that all characteristics "probably" arose multiple times by convergent evolution. Of course, just because we use Hennig's auxiliary principle we do not necessarily believe that convergences are rare or nonexistent. Convergences are facts of nature and rather common in some groups. But, in order to pinpoint convergence without invoking ad hoc assumptions, you must first have a tree, and without Hennig's auxiliary principle you will never get one. So back to the xylem and phloem: with Hennig's auxiliary principle you can deduce that plants that have xylem and phloem shared a common ancestor not shared with other plants. Of course, you don't make such a deduction in a vacuum. You "know" that "more primitive" plants lack these characters, and thus it is a good guess that the development of xylem and phloem is the derived state. This deduction is a primitive sort of outgroup comparison (we will discuss the ins and outs of outgroup comparisons in more detail later).

This principle represents the first step in Hennig argumentation. Basically, Hennig proposed that the phylogenetic puzzle should be solved by investigating individual characters and then combining the information from each character according to a set of rigorous rules.

> **Grouping Rule:** *Only synapomorphies (shared special homologies) provide evidence of common ancestry relationships. Symplesiomorphies (shared general homologies) and convergences and parallelisms (homoplasies) are useless in this quest.*

Convergent and parallel characters (both termed homoplasies) are useless indicators of common ancestry relationships because they evolved independently in each taxon that displays them (fig. 2.8a). The futility of using plesiomorphies in an attempt to reconstruct a particular phylogeny is more problematical. After all, plesiomorphies are homologies, so why can't they be used to seek common ancestry relationships? In fact, the answer is quite straightforward: depending upon the level of your analysis, these characters can be used because, *since evolution is an ongoing process, the plesiomorphic or apomorphic status of a character is a relative condition.* All plesiomorphies begin as evolutionary novelties (autapomorphies). So, a symplesio-

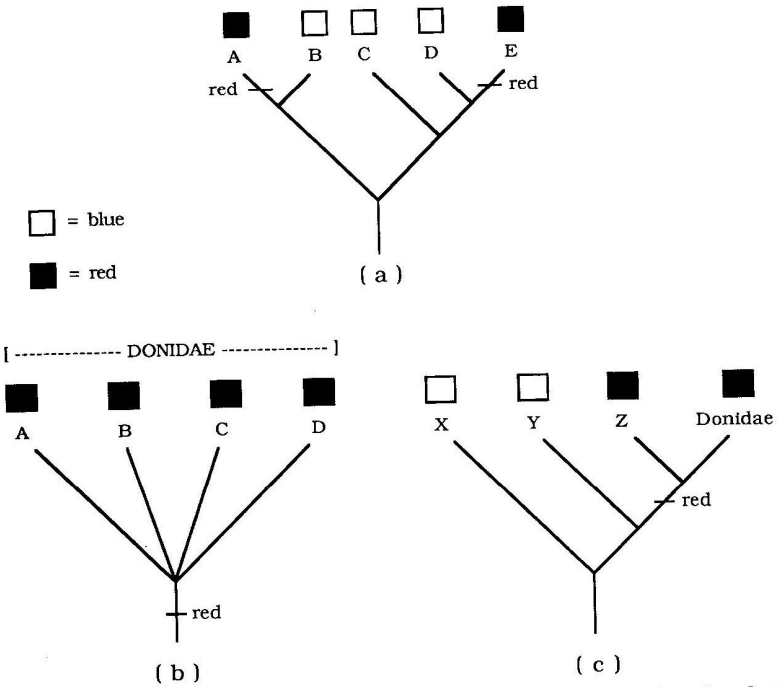


Fig. 2.8. The quest for phylogenetic relationships. *Letters* = taxa; *boxes* = the color of each taxon below. (*a*) Taxa A and E are both red, and other members of the group are blue; however, clustering A with E would be incorrect because, according to the phylogenetic tree based on numerous other characters, red arose independently in both taxa. (*b*) Since all members of the Donidae are red, that character does not tell us anything about individual relations among taxa A, B, C, and D (symplesiomorphy). (*c*) On a larger scale the Donidae share the character red with group Z (the tree is supported by many other characters); therefore red is useful in determining sister-group relationships at this level (synapomorphy).

morphy (character possessed by all members of the ingroup) cannot show common ancestry relationships within the group you are studying because it originated earlier than any of the taxa in your study group (fig. 2.8b). If you increase the temporal scale of your investigation, say by examining relationships among genera within one family instead of among species within one genus, this character will eventually prove to be useful (fig. 2.8c).

Finally, we have to consider how to combine the information from different transformation series into hypotheses of genealogical relationships. There are several ways to accomplish this. For now, we will use an old-fashioned (and perfectly valid) grouping rule that returns us to the roots of phylogenetic methods.

Inclusion/Exclusion Rule: *The information from two transformation series can be combined into a single hypothesis of relationship if that information allows for the complete inclusion or the complete exclusion of groups that were formed by the separate transformation series. Overlap of groupings leads to the generation of two or more hypotheses of relationship, since the information cannot be directly combined into a single hypothesis.*

The inclusion/exclusion rule is directly related to the concept of logical consistency. Trees that conform to the rule are logically consistent with each other, while trees that do not are logically inconsistent with each other. You can get an idea of how this rule works by studying the examples in figure 2.9. In figure 2.9a we have four characters and three potential trees. The first tree contains no character information, and since it provides no resolution of the phylogenetic relationships within the group, it is logically consistent by default with any tree that has character information. The second tree states that B, C, and D form a monophyletic group based on characters from two transformation series (1 and 2). The third tree states that C and D form a monophyletic group based on two additional characters (3 and 4). Note that the group C + D is completely included within the group B + C + D. Based on this distribution of characters, we hypothesize that the tree for these taxa includes C + D as a monophyletic group enclosed within a second monophyletic group B + C + D. Figure 2.9b shows the result of the inclusion of two monophyletic groups (A + B and C + D) within a larger monophyletic group (A–D). In this case, the groupings are consistent because A + B and C + D completely exclude each other. Finally, the example presented in figure 2.9c violates the inclusion/exclusion rule. Although both C + B and C + D can be included within the group B + C + D, the transformation series for characters 2 and 3 groups C + B and excludes taxon D, while the transformation series for characters 4 and 5 groups C + D and excludes taxon B. Thus, the phylogenetic information gleaned from these characters conflicts and the groups overlap (C is included in two different groups), producing two equally parsimonious trees that are locally inconsistent with each other.

The relationships of ABCidae

1. Transformation series (TS) 1 is composed of characters in the first column of the data matrix (table 2.1). Recall that plesiomorphies are coded as zero, synapomorphies are coded as one, and relationships among taxa are reconstructed based on shared derived traits or synapomorphies (the grouping rule). Given this, we can draw a tree with the groupings implied by the syn-

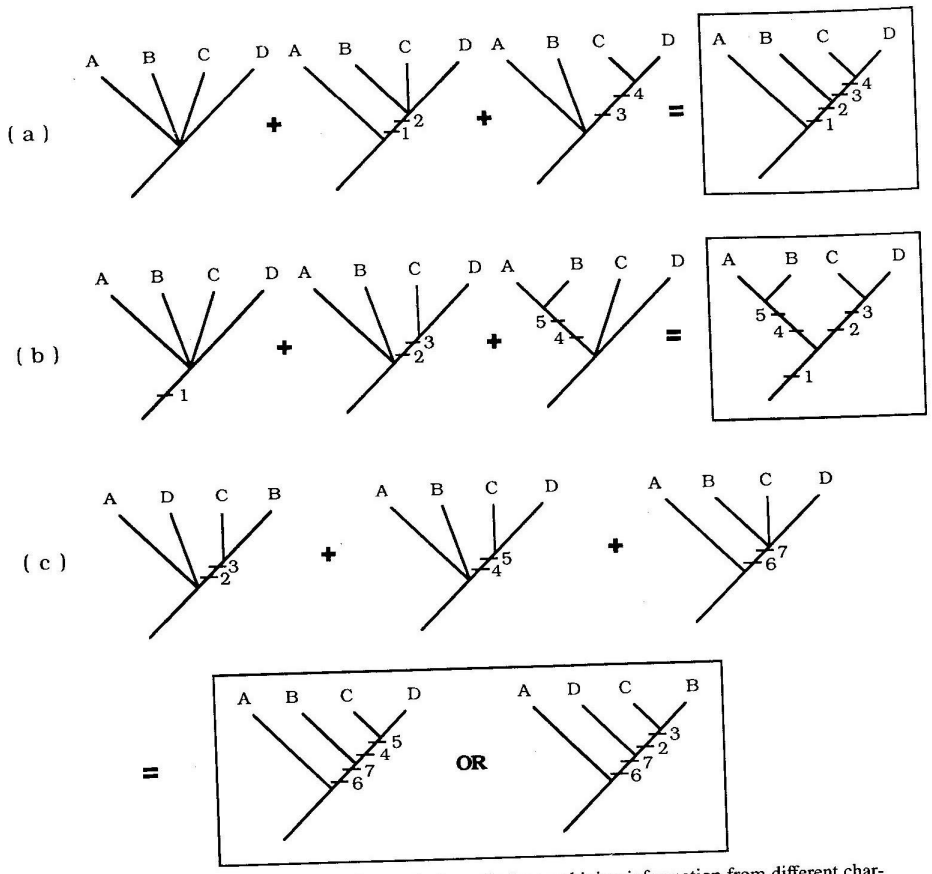


Fig. 2.9. The use of the inclusion/exclusion rule for combining information from different character transformation series into trees. (*a*) The group BCD completely includes the group CD. (*b*) The group ABCD completely includes the groups AB and CD, and, within this, the group CD completely excludes the group AB. (*c*) The group BCD completely includes the groups CD and CB, but groups CD and CB overlap because they both contain C. This results in the production of two different trees.

apomorphy found in the first transformation series (fig. 2.10a). We can then repeat the process for TS 2 (fig. 2.10b). Both trees, based on the distributions of two different characters, imply the same groupings; therefore, we can say that the trees are topologically identical, or isomorphic. Combining trees a and b according to the inclusion/exclusion rule produces the tree depicted in figure 2.10c (i.e., both characters support the group ABC to the complete exclusion of X). We can calculate a tree length for tree c by simply adding

Table 2.1 Data matrix for determining the relationships among taxa A, B, and C.

	Character Transformation Series						
Taxon	1	2	3	4	5	6	7
X (outgroup)	0	0	0	0	0	0	0
A	**1**	**1**	0	0	0	0	0
B	**1**	**1**	**1**	**1**	0	0	0
C	**1**	**1**	**1**	**1**	**1**	**1**	**1**

Notes: The matrix is composed of seven character transformation series and four taxa, the outgroup X, and the ingroup A + B + C. Synapomorphies are in bold type.

the number of synapomorphies that occur on it. In this case, the tree length is two steps.

2. Now, repeat this procedure for TS 3 and 4. Inspection of the data matrix reveals that the synapomorphies for these characters have identical distributions, implying that B and C form a monophyletic group (fig. 2.11a and b). If we combine the information from both characters, the results should look like the tree in figure 2.11c. This tree is also two steps long.

3. Only taxon C has the apomorphies listed in TS 5, 6, and 7. Apomorphic characters that are unique to one taxon are termed **autapomorphies.** Although they can tell us nothing about relationships among different taxa (fig. 2.12), such characters are useful diagnostic traits for identifying a particular taxon. For example, if we were to collect individuals displaying the autapomorphic state for characters 5, 6, and 7 (denoted by a one in the data matrix), we would assign those individuals to taxon C. On the other hand, collecting organisms bearing the synapomorphic condition for character 4 (also denoted by a one in the data matrix) only tells us that they are members of either taxon B or taxon C. Autapomorphies also count when figuring tree length, so the length of this tree is three steps.

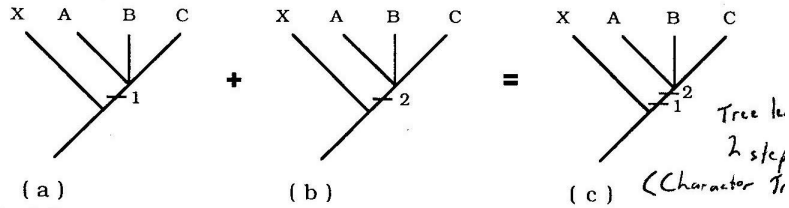


Fig. 2.10. Trees for the ABCidae, based on characters 1 and 2. (*a*) Tree produced by applying the grouping rule to character transformation series 1. (*b*) Tree produced by applying the grouping rule to character transformation series 2. (*c*) Tree produced by applying the inclusion/exclusion rule to the information provided by both characters.

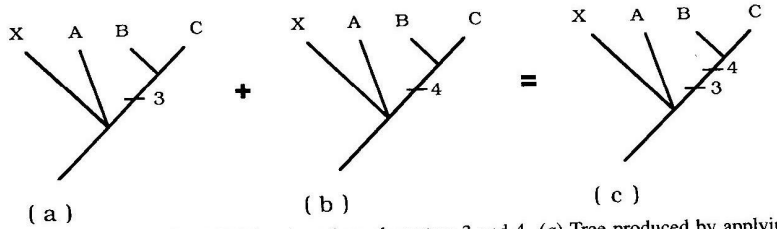


Fig. 2.11. Trees for the ABCidae, based on characters 3 and 4. (*a*) Tree produced by applying the grouping rule to character transformation series 3. (*b*) Tree produced by applying the grouping rule to character transformation series 4. (*c*) Tree produced by applying the inclusion/exclusion rule to the information provided by both characters.

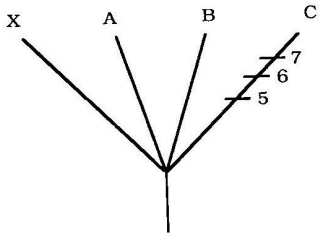


Fig. 2.12. Tree for the ABCidae, based on characters 5, 6, and 7. Since autapomorphies are not useful for grouping, this tree shows no resolution of relationships among the taxa.

4. We now have three different tree topologies (figs. 2.10c, 2.11c, and 2.12). If we examine these trees more closely, we discover that, although they are topologically different, they do not contain any conflicting information. For example, since TS 5–7 only imply that C is different from the other three taxa, this tree (fig. 2.12) does not conflict with the other two trees. Further, the distributions of TS 1 and 2 do not conflict with the distributions of TS 3 and 4 because TS 1 and 2 imply that A, B, and C form a monophyletic group, while TS 3 and 4 imply that B and C form a monophyletic group without saying anything about the relationships of A or the outgroup, X. Trees that contain different but mutually agreeable groupings are **logically compatible** or **fully congruent.** They can be combined without changing any hypothesis of homology, and when combined the length of the resulting tree is the sum of the lengths of each subtree. For example, all of the information in the data matrix can be combined to produce one tree (fig. 2.13) with a length of seven steps, exactly the number of subtree steps (2 + 2 + 3). This example produces a pattern similar to the one depicted in figure 2.9a: the group ABC completely includes the group BC and completely excludes the taxon X.

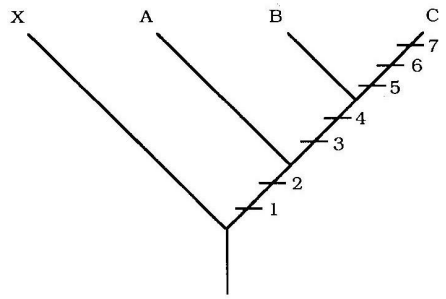


Fig. 2.13. Combining all the information in the data matrix (table 2.1) produces one hypothesis (tree) of the phylogenetic relationships within the ABCidae. This is the best estimate of the relationships, based on the available data. The tree proposes that taxa B and C are sister groups bound together by the possession of synapomorphies for characters 3 and 4, and that ABC is a monophyletic group based on the common shared characters 1 and 2.

The Relationships within the RSTidae

1. The data matrix for TS 1 and TS 2 (table 2.2) implies that R, S, and T form a monophyletic group (fig. 2.14).

2. TS 3 and TS 4 imply that S and T form a monophyletic group (fig. 2.15).

3. TS 5, 6, and 7 imply that S and R form a monophyletic group (fig. 2.16).

4. At this point you should suspect that something has gone wrong. TS 3 and 4 imply a monophyletic group that includes S and T but excludes R, while TS 5–7 imply a monophyletic group that includes R and S but excludes T. There must be a mistake, since we have violated the inclusion/exclusion rule. In such a situation we invoke the first principle of phylogenetic analysis: **there is only one true phylogeny.** Thus, one or more of our groupings must

Table 2.2 Data matrix for determining the relationships among taxa R, S, and T.

	Character Transformation Series						
Taxon	1	2	3	4	5	6	7
X (outgroup)	0	0	0	0	0	0	0
R	**1**	**1**	0	0	**1**	**1**	**1**
S	**1**	**1**	**1**	**1**	**1**	**1**	**1**
T	**1**	**1**	**1**	**1**	0	0	0

Notes: The matrix is composed of seven character transformation series and four taxa, the outgroup X, and the ingroup R + S + T. Synapomorphies are in bold type.

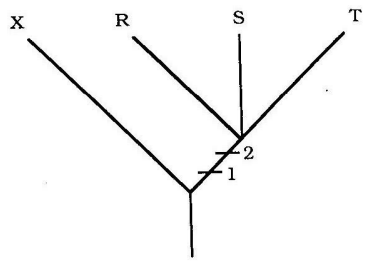


Fig. 2.14. Tree for the RSTidae, based on characters 1 and 2. This tree was produced by applying the grouping rule to character transformation series 1 and 2, then combining this information via the inclusion/exclusion rule.

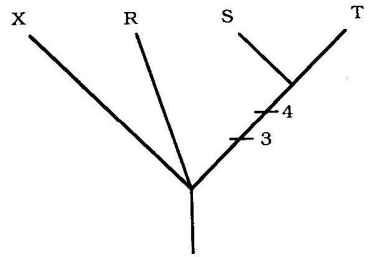


Fig. 2.15. Tree for the RSTidae, based on characters 3 and 4. This tree was produced by applying the grouping rule to character transformation series 3 and 4, then combining this information via the inclusion/exclusion rule.

be wrong. Fortunately the auxiliary principle keeps us going until we can *demonstrate* which of the groupings is incorrect. We are now faced with the problem of trying to differentiate between two logically incompatible trees (fig. 2.17). Note that there is some congruence between the trees based on their possession of the apomorphies from the first two TS (characters 1 and 2).

5. You have probably guessed by now that each of the trees in figure 2.17 is incomplete. Tree a lacks the transformation series 3 and 4, while tree b lacks the transformation series 5, 6, and 7. **Leaving characters out of an analysis is not acceptable.** In fact, eliminating characters that do not "fit" your hypothesis ranks among the top three heinous "crimes against phylogenetics" (the other two being grouping by symplesiomorphies and equating taxonomy with phylogeny). Adding the missing characters into both trees requires that we postulate that some of the evolutionary changes within this group are due to homoplasy (fig. 2.18). There are basically two types of homoplasy: a character may have risen independently more than one time (convergent or parallel character evolution), or there might be a reversal to

the "plesiomorphic" condition. We must consider both types of homoplasy in this example. Recall that our first tree (fig. 2.17a) neglected to include characters 3 and 4. There are potentially two ways to portray the distribution of these characters on the tree: either 3 and 4 arose independently in taxa S and T (fig. 2.18a), or 3 and 4 arose in the common ancestor of the group RST and were subsequently "lost" in taxon R (reversal to the ancestral [plesiomorphic] character conditions, fig. 2.18b). Examination of the distributions of characters 5, 6, and 7, which are missing on the second tree (fig. 2.17b), produces a similar pattern of homoplasy: either 5, 6, and 7 arose independently in taxa R and S (fig. 2.18c), or 5, 6, and 7 arose in the common ancestor of the group RST and taxon T subsequently reverted to the ancestral (plesiomorphic) character conditions (fig. 2.18d).

6. The question now becomes, Which of these trees should we accept?

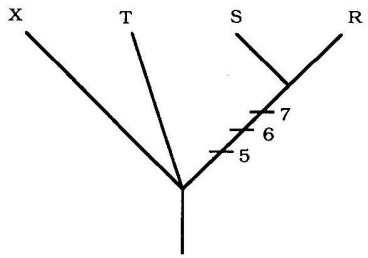


Fig. 2.16. Tree for the RSTidae, based on characters 5, 6, and 7. This tree was produced by applying the grouping rule to character transformation series 5, 6, and 7, then combining this information via the inclusion/exclusion rule.

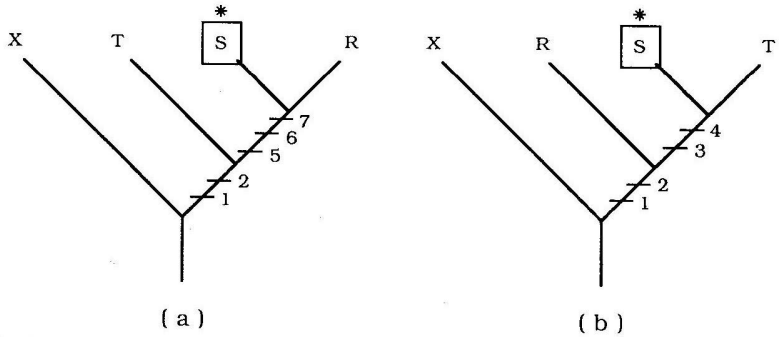


Fig. 2.17. Two logically incompatible trees produced from the information in the data matrix (table 2.2). Taxon S (marked with an asterisk) is the problem: characters 5, 6, and 7 place it with R, while characters 3 and 4 group it with T. Both trees cluster RST together based on possession of the apomorphic form of characters 1 and 2.

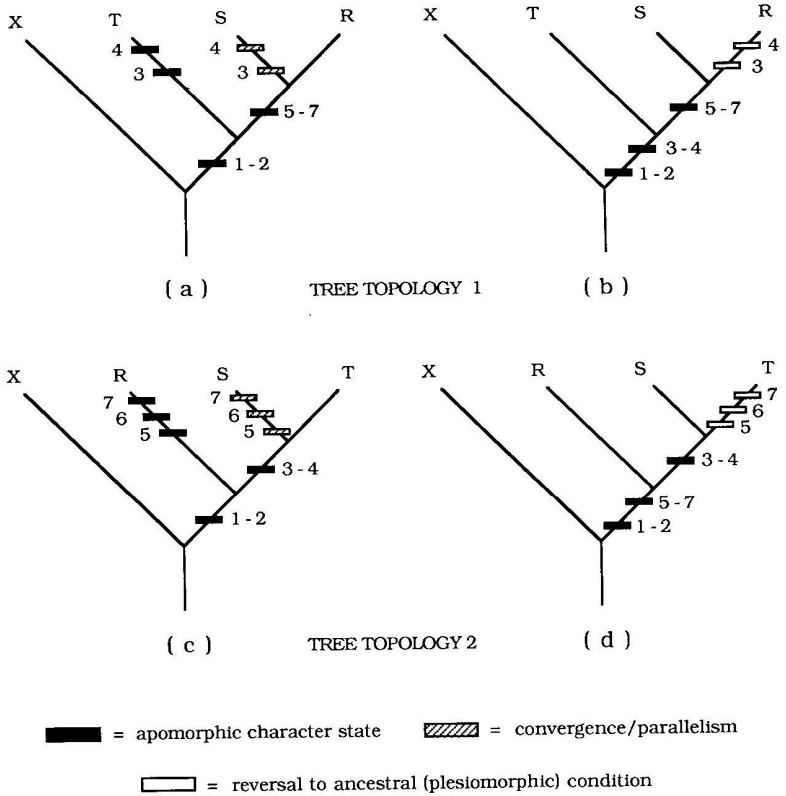


Fig. 2.18. Alternate hypotheses for the relationships of R, S, and T due to homoplasious characters. (*a*) Convergent evolution of 3 and 4 in taxa S and T. (*b*) Characters 3 and 4 revert to the plesiomorphic condition in taxon R. (*c*) Convergent evolution of 5, 6, and 7 in taxa S and R. (*d*) Characters 5, 6, and 7 revert to the plesiomorphic condition in taxon T.

That turns out to be a rather complicated question. If we adhere to the auxiliary principle, we should strive for a tree that includes the greatest number of homologies and the fewest number of homoplasies. Although these qualities are usually consistent with each other (i.e., the tree with the greatest number of synapomorphies is also the tree with the fewest number of homoplasies), you can find exceptions. Fortunately, numbers of homologies and homoplasies are related to tree length. Before we begin counting steps, notice that trees a and b (fig. 2.18) have the same topology; in fact, *these trees are the same hypothesis (tree) of phylogenetic relationships among the RSTid taxa even though they are based on different hypotheses of character change. This*

occurs because the topology of a tree is determined by synapomorphic relationships, not by the distributions of homoplasies. The same is true for trees c and d. So, as previously shown (fig. 2.17), we really have only two phylogenetic trees for this group. When you count the number of steps on each tree you discover that tree type 1 (fig. 2.18a and b) has nine steps while tree type 2 (fig. 2.18c and d) has ten steps. We accept tree 1 as the best estimate of phylogeny because it has a shorter length and thus incorporates the greatest number of homology statements and the fewest number of homoplasy statements for the data set. *Note that such statements are relative only to trees derived from the same data set*. The auxiliary principle coupled with the principle that there is only one phylogeny of life carried us to this point. Methodologically, we have employed the **principle of parsimony.** In the phylogenetic system the principle of parsimony is nearly synonymous with the auxiliary principle. At the moment, we cannot choose between the different sequences of character evolution postulated by trees a and b because both trees are equally parsimonious (the same length). This requires the use of criteria other than tree length. The examples we present in subsequent chapters of this book will not require such assessments. For those who are interested in these matters, we refer you to Wiley et al. (1990).

7. Finally, we can evaluate the performance of each character originally coded as a synapomorphy by calculating a **consistency index** (CI) for it. The CI of a character is simply the reciprocal of the number of times that character appears on the tree; therefore, true homologues (real synapomorphies) have a CI of 1.0. The CI (usually reported as a percentage) is a favorite summary "statistic" in computer programs such as PAUP and MacClade, so it is worthwhile practicing some hand calculations to remove some of the computer mystique! For example, in our most parsimonious tree (topology 1 in fig. 2.18), the apomorphy coded one in TS 3 appears twice on the tree, so CI = 1/2 = 50%. Based on this CI, we can see that this character is not really a synapomorphy. Of course, our best estimate of a true homologue is an a posteriori test because we can only calculate the CI after we have determined our best estimate of common ancestry relationships. There is no way to know in advance that a particular derived similarity will turn up with a CI less than 1. The interesting point about calculating consistency indices for individual characters is that, if you have the tree before you, it is completely unnecessary. Examination of the tree will tell you everything you need to know about the number of times a particular character has appeared during the evolution of your group. This calculation is simply a numerical convenience. More important to considerations of choices among multiple trees is the CI for the entire tree (Kluge and Farris 1969). Basically, this is a goodness-of-fit measure designed to indicate the degree of support for a particular tree by a

particular data set. The CI ranges from 0 to 100%, with a value of 100% indicating that all characters support the tree (no homoplasy). Once again, the calculation is very simple.

$$CI = \frac{\text{number of characters in the data matrix}}{\text{total number of characters on the tree}} \times 100$$

So, for the trees shown in figure 2.18, the consistency indices would be 7/9 = 77.8% (fig. 2.18a and b) and 7/10 = 70% (fig. 2.18c and d). For a given data set, trees that have the same length have the same CI. Tree length gives an indication of the quantity of character support, whereas the CI gives an indication of the quality of character support for a given tree.

Hennig (1966) and Brundin (1966) characterized the essence of phylogenetic analysis as the "search for the sister group." They recognized that if you could find the closest relative or close relatives of the group you are working on, you have the basic tools for deciding which characters are apomorphic and which are plesiomorphic in a transformation series. The argument goes something like this: You discover that members of a group have two different but homologous courtship characters, "zigzag dance" and "pummel dance." As a phylogeneticist you realize that one of these characters (the apomorphic one) might provide information about relationships within your study group, but that both cannot be equally informative because *you cannot group taxa based on plesiomorphic characters* (recall fig. 2.8b). If you find zigzagging in the sister group or in closely related groups (outgroups) of the taxon you are studying, then it is fairly clear that this dance type is older than pummeling; so zigzag dance must be the plesiomorphic character in the transformation series. The characteristics of members of related groups are thus vital components to decisions regarding the polarity of characters within the study group. The simplest rule for determining polarity is the

> **Relative Apomorphy Rule (outgroup comparison):** *Homologous characters found in the members of a monophyletic group and in the sister group are plesiomorphic, while homologous characters found only in the ingroup are apomorphic.*

Actual polarity decisions can be more complicated than our simple example. For example, what if (1) we don't know the exact sister group but have only an array of possible sister groups, (2) the sister group also has both characters, (3) either the ingroup or the sister group is not monophyletic, or (4) zigzagging evolved in the sister group independently? Answers to these questions depend on our ability to argue character polarities using some formal rules. We think the best discussion of these rules was published by Maddison, Donoghue, and Maddison (1984), although the issues have been discussed widely (see, e.g., Ross 1974; Crisci and Stuessy 1980; de Jong 1980; Stevens

1980; Watrous and Wheeler 1981; Wiley 1981, 1986a; Farris 1982; Patterson 1982; Donoghue and Cantino 1984; de Queiroz 1985). We will present the case developed by Maddison, Donoghue, and Maddison (1984) for groups in which sister-group relations have already been determined. Before proceeding with this discussion, however, we need to add some more terms to our phylogenetic vocabulary.

The ingroup is the group on which we are working. For purposes of character polarization arguments, the ancestor of the ingroup is depicted as an **ingroup node** (fig. 2.19a; note: the ingroup node = the ingroup, or ancestral, internode). *We are basically on a quest to determine what the character looked like in this ancestor because it represents the plesiomorphic (or ancestral) condition of the character for the ingroup.* The **outgroup node** is the node immediately below the ingroup node (fig. 2.19a). Characters are placed where taxa are usually labeled. Letters are used purely as a heuristic device to avoid connotations of "primitive" and "advanced." The ingroup is indicated by a polytomy since we presume that the relationships among these taxa are

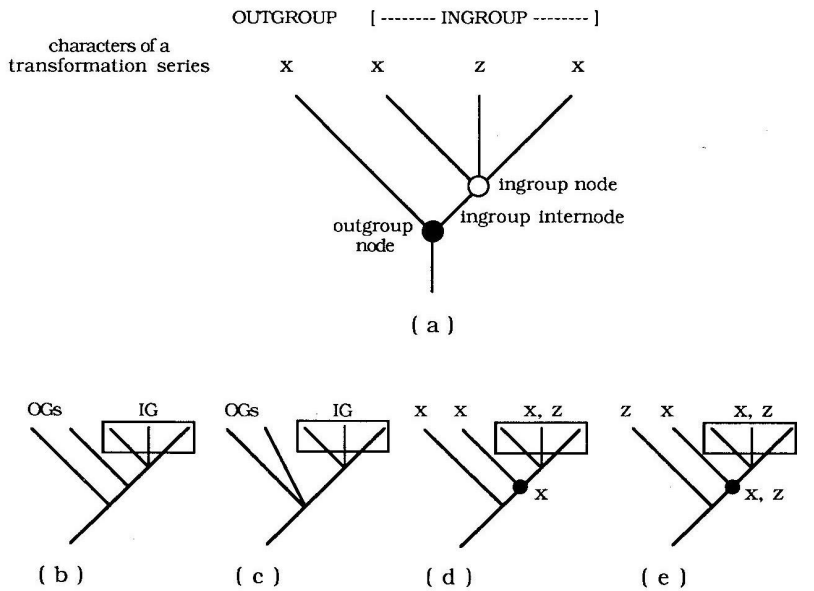


Fig. 2.19. Primer for character polarization. *Lowercase letters* = characters. Members of the ingroup (*IG*) are enclosed in a box. (*a*) Examples of some general terms used in character polarization. (*b*) Outgroup (*OG*) relationships are known. (*c*) Outgroup relationships are unknown. (*d*) State of the character at the outgroup node is known. (*e*) State of the character at the outgroup node is unknown.

Table 2.3 Matrix for characters found within the ingroup (Annidae), its sister group (the monophyletic group P + Q + R), and other outgroups.

	Taxon						
Character TS	M	N	O	P	Q	R	Annidae
1	b	a	a	b	b	a	a, b
2	b	b	a	b	b	a	a, b
3	a	b	b	b	b	a	a, b
4	a	a, b	a	b	b	a	a, b

Note: In this matrix the taxa are in columns and characters are in rows.

unknown. Note that this method applies only when the study group has both characters (otherwise there would be no need to polarize the character states). The relationships among outgroups can be either **resolved** (fig. 2.19b) or **unresolved** (fig. 2.19c). A decision regarding the character found at the outgroup node may be either **decisive,** which provides us with a best estimate of the condition found in the ancestor of our ingroup (fig. 2.19d: x is plesiomorphic and z is apomorphic in the ingroup), or **equivocal,** which does not allow us to postulate the direction of the character change (fig. 2.19e: we are not sure whether x or z is plesiomorphic).

Maddison, Donoghue, and Maddison (1984) began the quest for character polarities by attempting to determine the character state at the outgroup node. They reasoned that because this precedes the ingroup node, it will give us information about the character state in the common ancestor of the ingroup. Simple parsimony arguments are used in conjunction with an optimization routine developed by Maddison et al. (modified from earlier routines of Farris and Fitch). There are two cases: the relationships of the outgroups are known relative to the ingroup, and the relationships of the outgroups are either unknown or only partly resolved. The first case is the simplest, so we will begin by examining character evolution within the hypothetical taxon Annidae, its sister group P + Q + R, and other outgroups M, N, and O.

Character polarity in the group Annidae

1. Draw the phylogenetic tree of the ingroup and outgroups. We cannot reconstruct the tree on the basis of the characters in table 2.3, because these characters apply to the resolution of relationships within the ingroup, not to the relationships between the ingroup and the taxa in the outgroup. Presumably, then, we have access to a phylogeny for these taxa before beginning our investigation (fig. 2.20).

2. Inscribe each of the branches with its corresponding character from the first transformation series (i.e., the first row in the data matrix: fig. 2.21),

and indicate the six nodes. The lowest node on the tree is the **root node,** while the node connecting the Annidae to its sister group (PQR) is the **outgroup node.**

3. Label the two nodes, other than the root node, that are farthest from the outgroup node in the following manner: (1) Label the node "a" if the two closest nodes or branches are either both a or a and a, b. Notice that "closest nodes" refers to the nodes that are adjacent to the node in question, while "closest branch" is defined as any adjacent branch at the equivalent level or lower on the tree than the node in question. (2) Label the node "b" if the two closest nodes or branches are either both b or b and a, b. (3) If the closest branches/nodes have different labels (one a and the other b), label the node "a, b." Note that the lowest node, termed the **root node,** is not labeled. In

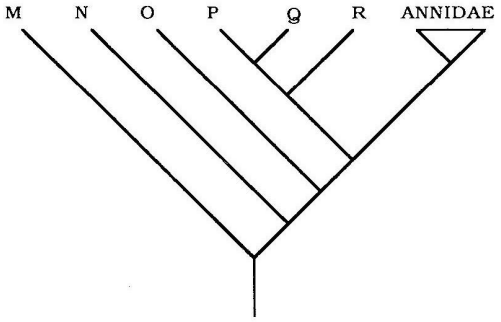


Fig. 2.20. Relationships of the Annidae clade to its closest relatives.

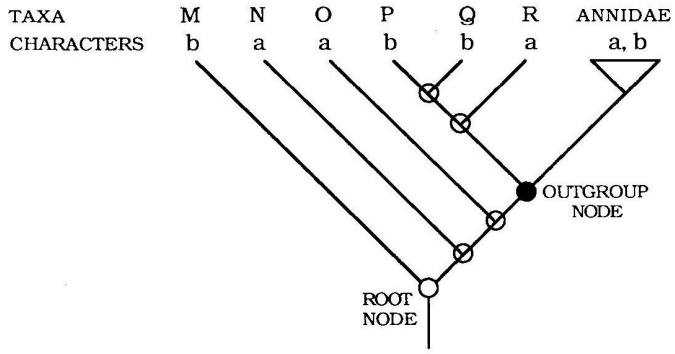


Fig. 2.21. Character states from the first transformation series in table 2.3, listed under the appropriate taxon. In order to determine the state of the character (either a or b) at the outgroup node, we must work towards it based on the information available at other, more-accessible nodes (depicted on this tree as *open circles*).

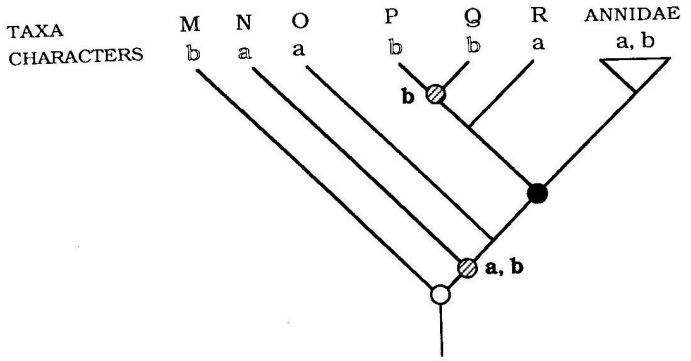


Fig. 2.22. First and second polarity decisions for character transformation series 1. *Outlined letters* = closest branches and nodes; *bold letters* = polarity decisions. The top node is labeled "b" because its closest branches are both b; the bottom node is labeled "a, b" because one of its closest branches is a and the other is b.

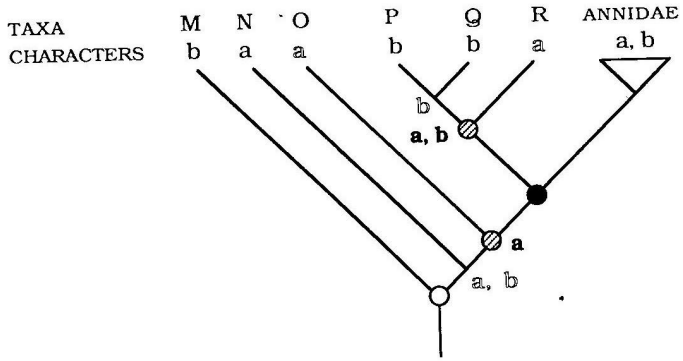


Fig. 2.23. Third and fourth polarity decisions for character transformation series 1. *Outlined letters* = closest branches and nodes; *bold letters* = polarity decisions. The top node is labeled "a, b" because its closest branch is a and its closest node is *b;* the bottom node is labeled "a" because its closest branch is a and its closest node is a, b.

order for us to label the root, we would need another outgroup. Since we are primarily interested in relationships within the ingroup, we will forget about this node. So, label the furthermost two nodes according to this procedure. The node connecting P + Q is given the notation b, because both P and Q have the character b. The node connecting N with O + P + Q + R + Annidae is given the notation a, b because M has character b and N has character a (fig. 2.22).

4. Continue working towards the outgroup node in the same manner (fig. 2.23).

5. The analysis is over when we reach a decision concerning the outgroup node. In this example, the assignment is a decisive **a** (fig. 2.24).

Repeating this procedure for character transformation series 2 of the matrix eventually produces an equivocal decision at the outgroup node (fig. 2.25).

One last thing. *Each of these decisions is made one transformation series at a time, and thus the polarization of every character occurs independently of every other character.* This does not mean that equivocal decisions based on single characters examined in vacuo will remain equivocal at the end of the analysis. The final disposition of character states is subject to an overall parsimony analysis which combines all of the information gleaned from each

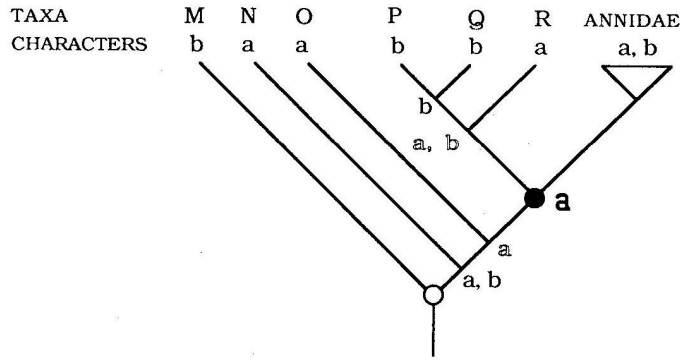


Fig. 2.24. Assignment of polarity to the outgroup node for character transformation series 1. *Outlined letters* = closest branches and nodes; *bold letter* = polarity decision. The outgroup node is labeled "a" because one of its closest nodes is a and the other is a, b.

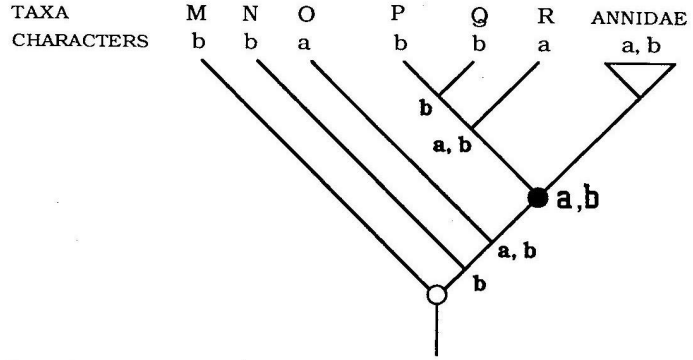


Fig. 2.25. Assignment of polarity to the outgroup node for character transformation series 2. The outgroup node is labeled "a, b" because both of its closest nodes are a, b.

of the separate character analyses according to the grouping and inclusion/exclusion rules.

Maddison, Donoghue, and Maddison (1984) discussed situations in which the relationships between the outgroups are either not resolved or only partly resolved. Since this watershed article is readily available, we will only mention two important observations: (1) Whatever the resolution of the outgroup relationships, the sister group is always dominant in its influence on the polarity decision. If the sister group is decisive for a particular state (i.e., a), no topology of outgroups further down the tree can result in a decisive b. (2) If you are faced with no sister group, but only an unresolved polytomy below the group you are working on, the frequency of a particular character among the outgroups in the polytomy has no effect on the decision at the outgroup node. For example, you could have ten possible, unresolved sister groups with character a and one with character b and the decision would still be equivocal at the outgroup node. Thus, *"common" does not equate with "plesiomorphic" in the phylogenetic system.*

Hennig argumentation was the original phylogenetic algorithm, and its application is still widespread. Working through the argumentation process for each character in a simple phylogenetic reconstruction is not overly difficult. But, even with relatively few taxa and characters, it is often laborious to reconstruct all of the character trees and the logically incompatible alternatives. We have good news and bad news about this situation. The good news is that computer algorithms have been designed to perform this laborious task (Kluge and Farris 1969; Farris 1970; Farris, Kluge, and Eckardt 1970). The bad news is that the ease with which computers perform this task may seduce us into forgetting that the results coming out are only as good as the data going in (Presch 1989). Increasing the amount of homoplasy in your data set increases the number of equally parsimonious pathways open to explain the evolutionary origins of these characters (i.e., increases the number of equally parsimonious trees). By contrast, increasing the number of homologous characters in an analysis decreases the number of equally parsimonious trees, because homologous characters correspond to a single phylogenetic tree. If the number of homologies has been greater than the number of covarying homoplasies in evolution, then the most biologically robust method of dealing with multiple equally parsimonious trees is to collect more data.

Character Coding for Building Trees

A **character code** is a numerical or alphabetical symbol that represents a particular character. We have already encountered codes in the preceding section. Using these characters and their codes has taught you something about the basics of tree reconstruction using classical Hennig argumentation

and some of the approaches to determining the polarity of characters through outgroup character argumentation. In this section you will be introduced to some of the different kinds of derived (apomorphic) characters encountered in phylogenetic research and some of the problems associated with assigning codes to these characters. *The ultimate goal here is to formulate rules for coding characters so they can be used to reconstruct phylogenetic relationships among taxa.*

Multistate Transformation Series

All of the derived characters we have considered thus far are qualitative characters and are parts of binary transformation series. A **binary transformation series** consists of a plesiomorphy and its single derived homologue. Binary transformation series present no problem in coding. You simply code each character according to the information available from outgroup argumentation and produce a matrix full of zeroes and ones. Complications can arise if you encounter **polymorphic taxa,** taxa that have both the plesiomorphic and the apomorphic characters; however, this problem is critical only when both characters are found in a single species. Considerable controversy surrounds the coding of such characters, especially when biochemical traits are used. There are two ways of handling this: code the taxon as displaying only the apomorphy and discount the plesiomorphy, because the plesiomorphic trait arose in an ancestor of the taxon (**qualitative coding**), or code according to the frequency of each character. There is considerable controversy about which of these two approaches is appropriate for phylogenetic analysis. (For excellent recent discussions of this problem, see Buth 1984, Swofford and Berlocher 1987, and Murphy 1988.)

Investigators working on a large group, or even a small group that has undergone considerable evolution, may discover that there are several different homologous characters in one transformation series. For example, if you were working on the phylogenetic relationships of fossil and recent horses, the transformation series for the number of toes on the hind foot would contain a goodly number of characters: four toes, three toes, and one toe in the ingroup, and five toes in the outgroup. Such a large transformation series, encompassing a plesiomorphic character and two or more apomorphic characters, is termed a **multistate transformation series.**

Multistate transformation series can be grouped according to the amount of information available concerning the pathway and direction of evolutionary change. As discussed earlier, this produces four possible types of multistate transformation series (see fig. 2.6): ordered, unpolarized; unordered, unpolarized; ordered, polarized; and unordered, polarized. Binary characters are much simpler because they are all automatically ordered (but not neces-

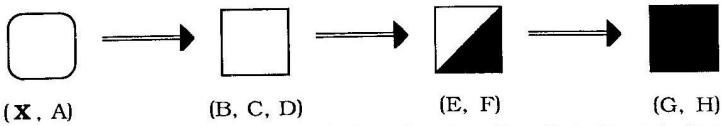


Fig. 2.26. A simple linear character tree for four characters. Taxa displaying each character are listed in parentheses below the character. X = outgroup; A–H = ingroup taxa.

polarized). **A linear transformation series,** consisting of characters related to one another in a straight-line fashion, is the simplest multistate transformation series (fig. 2.26). The relationships of these characters can be termed a **character tree,** and in this particular case you can see that there are no branches on the tree. Notice that this character tree is both ordered and polarized. At the moment we are not concerned with how to reconstruct the character tree, but rather with how to code the information presented in this tree for use in reconstructing a phylogenetic tree (we will return to the question of just how these transformation series are derived at the end of this section). ***It is important to understand that a character "phylogeny" is not the same as a phylogeny of taxa. A character tree only contains information about the relationships among characters. It is not a hypothesis of the underlying phylogenetic relationships of the taxa.*** This is a critical, and often misunderstood, distinction.

A simple linear transformation series

The data matrix for the linear transformation series (fig. 2.26) can be coded in one of two basic ways. First, we can simply code the characters in a linear fashion, assigning a value to each character based on its position in the sequence. For example, we have chosen to code rounded square as zero, square as one, black-and-white square as two, and black square as three. Each value is placed in the data matrix in a single column (table 2.4). Every apomorphy will contribute to the length of our forthcoming phylogenetic tree in an additive fashion. We use the term "additive" because each instance of evolutionary modification requires one step along the tree and counting all of the steps in a straight line shows exactly how much the transformation series added to the overall tree length. Indeed, such transformation series are often termed additive multistate characters.

We could also use **additive binary coding,** a method that breaks the character down into a number of binary subcharacters that are each represented by their own column of information. For example, we can consider both black-and-white square and black square as subsets of white square, since each is ultimately derived from white square. The first additive binary column in the data matrix reflects this fact, coding rounded square as the plesiomorphic character (0) and white square plus all of its descendants as apomorphic (1). Now, black square is a subset of black-and-white square. The second additive binary column contains the codes for this next level of comparison: both rounded square and white square are plesiomorphic relative to black-and-white square, so they receive a zero, while the black-and-white square and its descendant black square receive a one. Finally, on the last level, black square is apomorphic relative to all the other characters in this linear series, so it is coded one in the third column and everything else receives a zero. Three columns now represent the transformation series. You can double-check your binary coding by adding all the ones together in each row and placing them in a single column. You should find that you have replicated the original linear coding column. Either method of coding produces exactly the same character sequence and thus contributes the same information to a phylogenetic reconstruction (fig. 2.27).

If you choose additive binary coding, it is important to remember that you are dealing with one homologous character series, and not three independent characters. If you count each state as an independent character, you may artificially inflate the consistency index and overestimate the degree of support for your tree. So bear this in mind if you are using a computer program to analyze your data, because computers cannot differentiate between a single multistate transformation series and independent binary transformation series for each of the derived states in the multistate transformation series.

Table 2.4 Data matrix for the four characters and nine taxa in the linear transformation series in figure 2.26.

Taxon	Linear Coding	Additive Binary Coding		
X (outgroup)	0	0	0	0
A	0	0	0	0
B	1	1	0	0
C	1	1	0	0
D	1	1	0	0
E	2	1	1	0
F	2	1	1	0
G	3	1	1	1
H	3	1	1	1

A branching transformation series

A **branching transformation series** (nonadditive or complex transformation series) contains characters that are related to each other in a branching, rather than a straight-line, fashion (fig. 2.28). Notice that this character tree is both ordered and polarized. Since the characters are not related in a linear

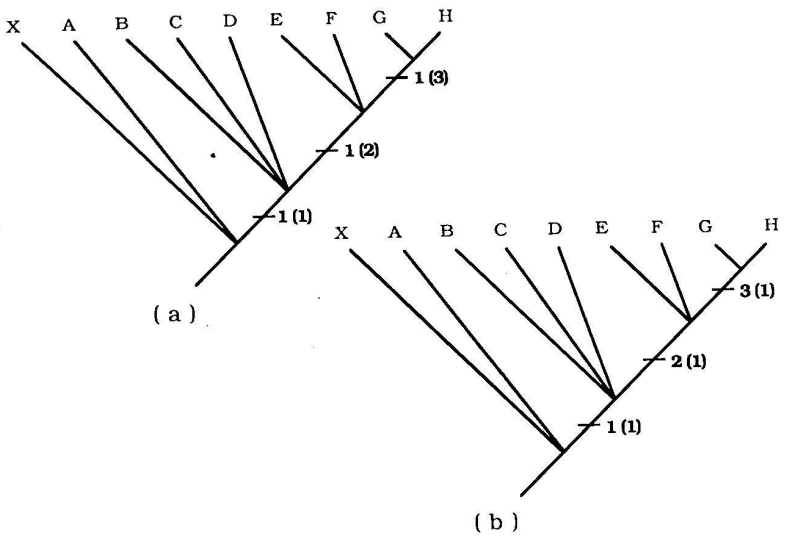


Fig. 2.27. Phylogenetic tree constructed from the linear character tree in figure 2.26. (*a*) Simple linear coding breaks the character into four parts, the plesiomorphic state (0), which need not be represented on the tree, and three apomorphic states (1, 2, and 3; one character, four states). (*b*) Additive binary coding breaks the linear series into three independent characters, then groups according to the apomorphic states (1) of those characters (three characters, two states each). Both methods of coding preserve the phylogenetic information in the character intact and thus will produce the same phylogenetic tree.

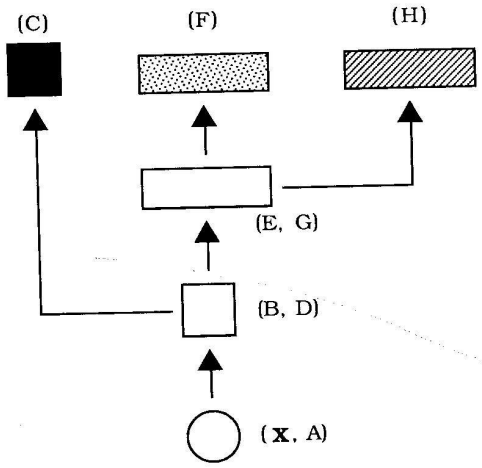


Fig. 2.28. A complex branching character tree for six characters. Taxa displaying each character are listed in parentheses.

fashion, simple additive coding will result in errors in translating the transformation series into a phylogenetic tree.

Labeling the characters of a branching transformation series in a simple linear fashion will obviously result in some misinformation, so how can we display these complex relationships? There are two basic methods: additive binary coding and mixed coding. Since we are already proficient at binary coding, let's turn to this method first.

We arrived at the additive binary codings in the data matrix (table 2.5) by working through the following steps. First, code the first additive binary column: square, which is ancestral to all other characters in the character tree, is apomorphic relative to circle; therefore, square and all of its descendants are coded as apomorphic (1), while circle is coded as plesiomorphic (0). The relationship between square and circle is a simple linear one. Moving to the next comparison, the relationship between square and its most immediate modification brings us to a branch point on the tree, because both black square and white rectangle are directly derived from square. We deal with branch points by examining the derived states one at a time. Second, code the second additive binary column: black square is derived directly from square; therefore square is coded as plesiomorphic (0) and black square as apomorphic (1). Notice that *only the taxa displaying either the apomorphic character or any of its modifications are assigned a code of one; all other taxa receive a zero*. In this case, black square has no modifications and is only present in taxon C (it is an autapomorphy). Third, code the third additive binary column: rectangle is derived directly from square; therefore square is coded as zero and rectangle as one. Again, the taxa bearing either the apomorphic condition or any of its modifications are assigned a code of one, so taxa E and G (apomorphic state rectangle) and F and H (bearing the descendants of rectangle) receive a one; all other taxa receive a zero. Fourth, code the fourth additive binary column: speckled rectangle is directly derived from rectangle and is an autapomorphy for taxon F; therefore F is assigned a one

Table 2.5 Data matrix for the six characters and nine taxa in the branching transformation series in figure 2.28.

Taxon	Additive Binary Coding					Mixed Coding		
X	0	0	0	0	0	0	0	0
A	0	0	0	0	0	0	0	0
B	1	0	0	0	0	1	0	0
C	1	1	0	0	0	1	1	0
D	1	0	0	0	0	1	0	0
E	1	0	1	0	0	2	0	0
F	1	0	1	1	0	3	0	0
G	1	0	1	0	0	2	0	0
H	1	0	1	0	1	2	0	1

for this character and everyone else gets a zero. Finally, code the fifth additive column: striped rectangle is directly derived from rectangle and is an autapomorphy for taxon H; therefore, only H receives a one for this character.

Mixed coding, also called **nonredundant linear coding,** is a hybrid between additive binary coding and linear coding. By convention, the longest straight-line branch of the character tree is coded in a linear fashion, then branches off this linear tree are coded in an additive binary fashion. Depending on the asymmetry of the character tree, this strategy can substantially reduce the number of character columns. Returning to the character tree, the longest sequence of character modifications is circle, square, rectangle, speckled rectangle (note: we could also have chosen the sequence circle, square, rectangle, striped rectangle; however, since nodes can be freely rotated on a tree, the choice between the "sister characters" speckled rectangle or striped rectangle is completely arbitrary and does not change the outcome of the analysis). We place the codes for this section of the tree (0, 1, 2, 3) in the first mixed coding column of the data matrix (table 2.5). Branches from this sequence, the autapomorphies black square and striped rectangle, are then each assigned a code of one in a separate column as described in the preceding section.

Polarization Arguments and Multistate Transformation Series

As promised above, we will now return to the methods involved in determining the evolutionary sequence of multistate characters. Sometimes the ordering of these transformation series can be determined using biological data, such as information about developmental sequences (Nelson 1978; Nelson and Platnick 1981; Patterson 1982; Voorzanger and van der Steen 1982; Brooks and Wiley 1985; de Queiroz 1985; Kluge 1985) or directions of biochemical pathways (Seaman and Funk 1983). More often, though, these data are not available or are not reliable.

What can phylogenetic systematic analysis tell us in the absence of this information? For example, suppose we wish to examine a group of mammals that includes species that do and do not have forelimbs modified into wings. Furthermore, extensive research has revealed that there are two types of wings (let us say short and long, just for this example). Using all other mammals as outgroups reveals that "no wings" came first (is plesiomorphic), but from that point, the evolutionary pathway of wings could have been (1) short to long, (2) long to short, or (3) short and long arising independently from the "no wings" condition. For these species, simple outgroup comparison does not resolve the polarity of the apomorphic states because they do not occur among members of the original outgroups. In order to resolve the transformation series for apomorphic states restricted to the ingroup, it is necessary to find outgroups that possess at least one of these states. In such cases,

phylogeneticists determine character polarities by using a method that preserves the logic of outgroup comparisons. This method, developed by Watrous and Wheeler (1981), is called **functional outgroup analysis.**

Let us begin with an ingroup (taxa A-E), a set of outgroups, and three characters (table 2.6). One binary character supports the monophyly of the ingroup. Two characters, one binary and the other multistate, help us resolve relationships *within* the ingroup.

Table 2.6 Distribution of states for one multistate and two binary characters among five members of an ingroup (A–E) and a set of outgroups.

	Characters		
Taxon	Binary		Multistate
Outgroups	0	0	x
A	1	0	x
B	1	0	y
C	1	+	y
D	1	+	z
E	1	+	z

Note: The first binary character supports the monophyly of the ingroup.

1. Use the second binary character to produce a partial phylogenetic tree for the ingroup. Figure 2.29 depicts the distribution of states for this character among the outgroup and ingroup taxa.

2. State 0 is clearly supported as the plesiomorphic condition by outgroup comparison. Hence, taxa C + D + E are united as a group within the ingroup by the shared possession of the derived (synapomorphic) state + (fig. 2.30). Because C + D + E is a putative clade, we can consider it to be a "functional ingroup." A and B then form the "functional outgroups," because their logical relationship to C + D + E is the same as the relationship of the outgroups to A + B + C + D + E.

3. Use the original outgroups and the functional outgroup to polarize the multistate transformation series. First, place the corresponding character states x, y, and z at the tips of the branches for the outgroup and ingroup taxa (fig. 2.30). Polarization via the "outgroup" taxa indicates that character state x is plesiomorphic to either y or z. Now, examine the functional outgroups. One of them (A) exhibits state x, while the other (B) exhibits state y. One of the functional ingroup members (C) exhibits state y, while the other two (D and E) exhibit state z. Because state y occurs in **both** the functional ingroup and the functional outgroup, we conclude that state y is plesiomorphic to state z. The original outgroups tell us that x arose first, while the functional outgroup tells us that y arose next. Consequently, state z arose last, and the

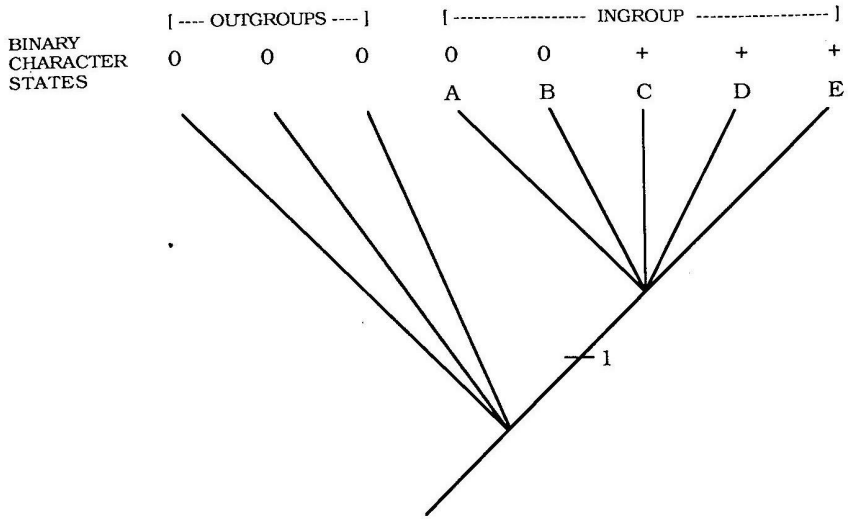


Fig. 2.29. Starting point for phylogenetic analysis of ingroup A–E using three outgroup taxa. Synapomorphy 1 supports the monophyly of the ingroup. Distribution of states (0 and +) for the binary character are indicated at the tips of each branch.

transformation series for this multistate character is the linear sequence x to y to z.

4. Combine the information from the binary and multistate characters. Outgroup comparisons for the binary character support the interpretation that state 0 is plesiomorphic and state + is apomorphic, linking taxa C + D + E as a clade. Functional outgroup analysis suggests that state x is plesiomorphic for the ingroup, that state y is apomorphic to x, linking B + C + D + E as a clade, and that state z is apomorphic to y, linking D + E as a clade. Invoking the inclusion/exclusion rule results in a fully resolved phylogenetic tree (fig. 2.31).

To return to our example of winged mammals, we use some characters to tell us that one group of mammals (ingroup) is distinct from other mammals (outgroup). We then use other characters to give us enough divisions among the winged members of this ingroup (functional in- and outgroups) to establish the sequence in which short and long wings evolved.

The use of functional outgroup comparisons (the logical justification of which is called "reciprocal illumination": Hennig 1966; Wiley 1981) strikes many people as being circular. Although the distinction is a fine one, we believe that the curse of circulatory is avoided because *the groupings within the original ingroup that allow us to determine polarities for the multistate character are determined a priori by reference to other traits.* Of course, if we do not have very many other characters, the robustness of our assessment

of multistate polarities is not very great. Consequently, many phylogeneticists worry about using multistate characters at all. In a complementary manner, other phylogeneticists worry about problems associated with treating states of an evolutionary sequence as independent characters. For example, if we treated x, y, and z as three binary characters, we would be able to use the original outgroups to determine that "x present," "y absent," and "z absent" were plesiomorphic conditions. The states "y present" and "z present" would be interpreted as synapomorphies of B + C and of D + E, respectively. Considering "z present" as a synapomorphy of D + E would not conflict with other data (i.e., the second binary character), but considering "y present" to be a synapomorphy of B + C would conflict. In this case, breaking x, y, and z into independent characters would result in unnecessary postulates of homoplasy (violating Hennig's auxiliary principle). This is the reason phylogeneticists study each character carefully, argue transformation series as independently as possible, often considering both binary and multistate options, and constantly seek additional data in an effort to provide the best possible hypothesis at any given time.

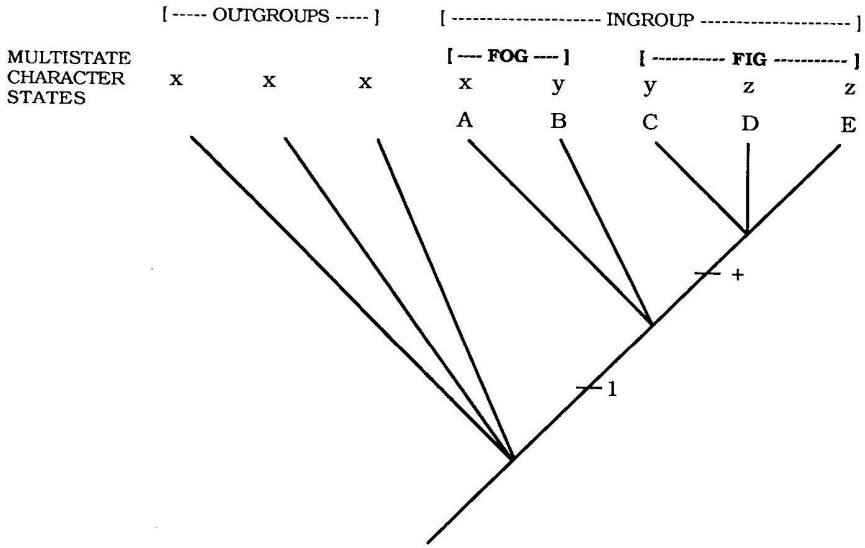


Fig. 2.30. Partial resolution of phylogenetic relationships among members of ingroup A–E, based on distribution of states for one binary character. Outgroup comparisons support the interpretation that state + is apomorphic, linking taxa C + D + E into a clade within the ingroup. This clade now serves as a functional ingroup (*FIG*) within the original ingroup, with the other members of the original ingroup (A and B) serving as functional outgroups (*FOG*). The distribution of states for a multistate character (x, y, and z) among members of the ingroup and outgroup is indicated above the taxa.

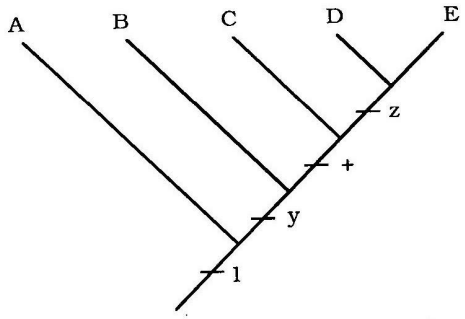


Fig. 2.31. Phylogenetic tree for taxa A–E, based on outgroup comparisons for two binary characters, and outgroup and functional outgroup comparisons for one multistate character.

Summary

Operationally, phylogenetic systematics can be summarized in four steps. First, use nonphylogenetic criteria to assess similarities in traits among species and assume homology whenever possible. Second, use outgroup comparisons to distinguish shared general, shared special, and unique homologies. Third, group according to shared special homologies, mapping all traits onto the resultant tree to show the distribution of the total data base. Fourth, interpret inconsistent traits, post hoc, as instances of homoplasy (parallel and convergent evolution); that is, as cases in which the initial nonphylogenetic homology criteria were misleading. The result of this procedure is a hypothesis of the phylogenetic relationships within the study group.

The number of branching diagrams appearing in the evolutionary biology literature is growing. Many of these are called "cladograms"; however, not all of these diagrams are constructed using phylogenetic systematic methods. The information presented in this chapter will help you critically evaluate these diagrams, for *only trees produced in accordance with phylogenetic systematic principles provide the robust estimates of genealogical relationships that are the necessary precursors for historical ecological studies.*

Answers to Some Common Questions and Misconceptions

What happens if you change the outgroup?

The outgroup's function is to identify the plesiomorphic character state for characters in the ingroup. So changing the outgroup could potentially cause one of the following things to happen to your tree.

1. There might be no change. This occurs when all the outgroups share with the ingroup the same plesiomorphic state for a given character.
2. There might be a change in the polarity of some transformation series.

This would happen if the outgroup and the ingroup share *numerous* homoplasious, apomorphic characters. In this case you look for a consensus state among outgroups and hope that you have enough other characters to provide a tree that highlights the homoplasy. If you do not, you will get multiple trees, and this, in turn, indicates that you need to collect more data.

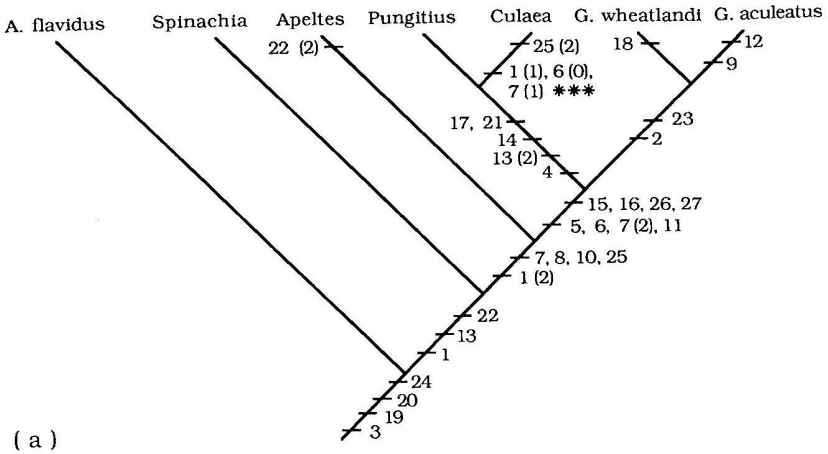
3. There might be a loss of resolution. If the new outgroup is distantly related to the ingroup, the groups might not share enough similarities for comparison, and this will lead to an increase in the number of unresolved polytomies in the tree. This produces an incomplete, rather than an "incorrect," estimate of phylogeny.

That changing outgroups will change the phylogenetic tree is an important and common misconception about phylogenetic systematics. Figure 2.32 shows a series of trees based on twenty-seven behavioral characters for the gasterosteid fishes (sticklebacks). Characters were polarized in the following manner: first, using the sister group of the Gasterosteidae, the Aulorhynchidae, or tubesnouts (fig. 2.32a); second, using a more distantly related species, the pipefish. *Sygnathus typhle* (fig. 2.32b); and third, using a very distantly related species, the salmonid *Onchorhynchus nerka* (fig. 2.32c). Interestingly, although all outgroups polarized some characters differently, the three tree topologies are consistent with one another; in fact, the trees based on the tubesnouts and the salmonid are identical. Although the phylogenetic relationships are retained, the number of informative characters decreases with use of the other outgroups. Overall, then, the distinction is not one of "correct" versus "incorrect" but rather one of the relative degrees of information available from the particular outgroup chosen. The more robust a tree proves to be in response to changing the outgroup, the greater confidence is instilled in it.

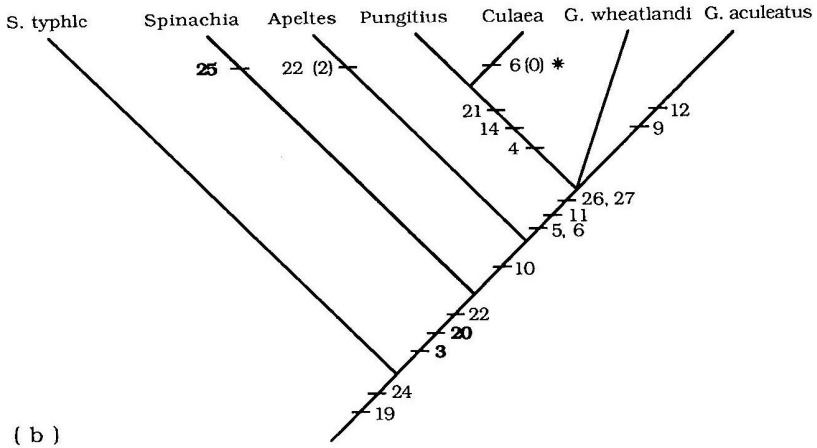
By now you are probably asking, "But even if you don't change the tree topology, haven't you changed some of the character transformation hypotheses by using different outgroups?" This is a critical question because, as the distribution of traits in figure 2.32 demonstrates, changing outgroups can change the hypothesized transformation series for some characters, and this, in turn, will affect our evolutionary hypotheses for those characters. Because of this, it is important to follow one cardinal rule: *never use the characters that are part of the evolutionary hypothesis under investigation to build your phylogenetic tree.* Rather, these characters should be mapped onto an existing tree. We will discuss methods for mapping characters (optimization) in chapter 5.

Is outgroup comparison an exercise in circular reasoning?

No. You do not have to have the sister group to polarize characters, so you do not have to have a phylogenetic scheme a priori in order to determine the



(a)



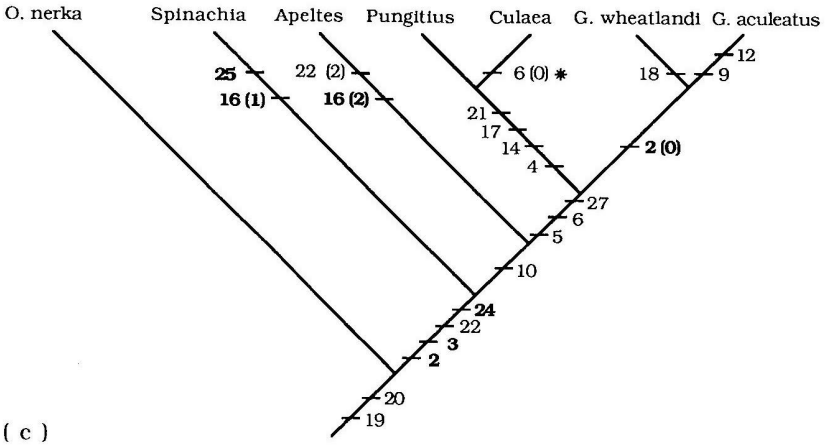
(b)

outgroups, and the polarizations for each character are argued independently based upon comparisons with the outgroups.

Does the method assume or rely on any particular evolutionary mechanism?

No. It assumes only that evolution has occurred. In fact, the utility of phylogenetic systematics in evolutionary biology stems from its independence of any particular evolutionary mechanism.

Fig 2.32 *(continued)*



(c)

Fig. 2.32. What happens if you use a different outgroup? (*a*) Polarizations using the sister group, the tubesnout *Aulorhynchus flavidus*. (*b*) Polarizations using a member from the next most closely related order, the pipefish *Sygnathus typhle*. Characters in *bold type* are polarized differently using this outgroup. Characters 1, 2, 7, 8, 13, 15, 16, 17, 18, and 23 cannot be polarized using this outgroup; therefore, they are eliminated from the analysis. Note that although there is a loss of resolution in the polytomy among *Pungitius/Culaea* + *Gasterosteus wheatlandi* + *G. aculeatus*, the overall topology of this tree is compatible with the first tree. (*c*) Polarizations using a distantly related teleost, the salmonid *Onchorhynchus nerka*. Characters 1, 7, 8, 11, 13, 15, 23, and 26 cannot be polarized. This tree is identical to the first tree. * = homoplasious characters; *numbers in parentheses* = the state of a multistate character.

Isn't parsimony an assumption of evolutionary mechanism?

Parsimony is a scientific principle used by scientists to make decisions about ambiguous data. Basically this principle can be stated in the following manner: when there are conflicting hypotheses for a given data set, accept the hypothesis that is supported by the greatest amount of data. The use of parsimony in phylogenetic systematics is no different from its use in any other

branch of biology or any other science, does not invoke any particular evolutionary mechanism, and does not mean that systematists believe that evolution is parsimonious or that "parsimony" equals "truth." Invocation of this principle simply gives us a starting point for comparative studies. From there, any author who prefers a less parsimonious tree must justify this choice by providing corroborating *biological* evidence.

Are phylogenetic trees tests of evolutionary mechanisms?

Phylogenetic trees are *not* tests of evolutionary mechanisms, they are descriptions about patterns in nature. From these patterns we can obtain critical evidence about some evolutionary principles, which, in turn, may help us to design an experiment to test the existence of a hypothesized mechanism.

When do you have enough characters?

The reconstruction of phylogeny is an open-ended process, so in principle you never have enough characters. In practice, you stop when you stop getting different answers or different resolutions when you add data. Even then it is possible for someone else to come along and modify what you have done.

Is there any inherently preferred data type?

No. It is expected that any data, analyzed phylogenetically, will give the same or highly similar answers because organisms are the result of a dynamic interaction among genetic, developmental, physiological, morphological, and behavioral systems through time. There is no a priori reason to believe that one system contains more information about evolutionary relationships than any other.

If you rotate the branches about a node on a phylogenetic tree, do you change the implied phylogenetic relationships?

No.

What's wrong with equating a taxonomic classification with phylogeny?

It is inappropriate to use a taxonomic classification as a phylogeny because many classifications portray paraphyletic (or polyphyletic) taxa as monophyletic groups. Evolutionary explanations based on such classifications will overestimate the importance of adaptive plasticity, because diagnoses for paraphyletic groups list synapomorphic traits more than once. This gives the impression that these traits are actually examples of parallel or convergent

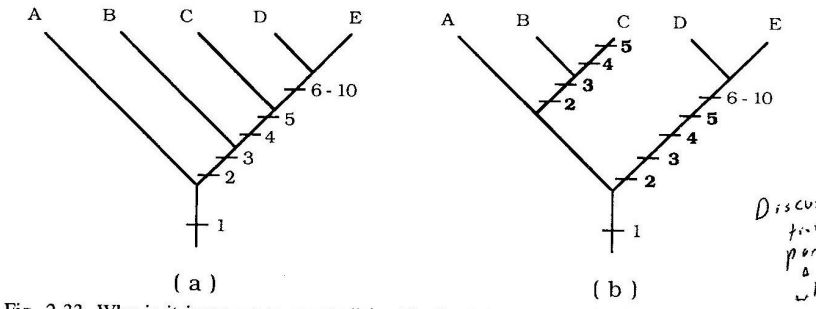


Fig. 2.33. Why is it incorrect to equate "classification" (or taxonomy) with "phylogeny"? (*a*) Phylogenetic tree reconstructed for species A–E based upon phylogenetic systematic methods. (*b*) Tree reconstructed from a taxonomic classification scheme that includes the paraphyletic group A + B + C. This forces us to postulate that characters 2, 3, 4, and 5 evolved twice.

evolution, and such homoplasy, in turn, is often considered strong evidence of adaptive evolution.

Consider the following example. Figure 2.33a depicts the phylogenetic tree for a group of hypothetical taxa. Although the presence of five synapomorphic traits (characters 6–10) distinguishes species D and E from species A, B, and C, these taxa are all members of a monophyletic group characterized by the presence of traits 1–5. Now suppose we have a classification scheme that places species D and E in one taxon because they are so distinct, and places species A, B, and C in another taxon. Reconstruction of phylogenetic relationships based on this classification will produce the tree in figure 2.33b. This arrangement forces us to postulate that characters 2–5 evolved twice, overestimating the amount of adaptive evolution. Since most commonly accepted classifications include paraphyletic groups, they cannot serve as independent templates for estimating the origins, elaborations, and associations of ecological characters through evolutionary time. Unfortunately, given the current dearth of available phylogenetic trees, many researchers have been forced to utilize classifications in their preliminary analyses of behavioral/ecological evolution.

Cladograms only represent branching points in evolution and cannot represent the relative degree of evolutionary divergence among lineages, can they?

If we have a large sample of characters, it is possible that the relative degree of evolutionary divergence among lineages can be estimated. This could be obtained by comparing the relative numbers of apomorphic transformations between sister groups on the phylogenetic tree. What we lack at

the present are protocols for letting us know when we have a proper sample of characters.

Why aren't there primitive and advanced species, just like there are primitive and advanced characters?

Because not all characters evolve at the same rate and to the same degree in different lineages. As a consequence, all species are mosaics of plesiomorphic and apomorphic traits, and it is inappropriate to speak of plesiomorphic and apomorphic species. One can speak of sister-group relationships or of relative position in a phylogenetic tree.

Can't you just manipulate the data to get the answer you "want"?

Of course, but this problem is one of scientific ethics and not unique to either phylogenetic systematics in particular, or biology in general. In fact, phylogenetics may be slightly more open to scrutiny because you have to report both the character descriptions and codings in each publication; therefore, the original data are more accessible to reanalysis than, say, a table of p values.