

Chapter 8

Developing Classifications

Systematics aims at developing classifications based on different criteria and, often a distinct methodology is employed for the analysis of data. Data handling to establish relationships between the organisms often makes use of one of the two methods: **phenetic methods** and **phylogenetic methods**, often providing different types of classification. Distinction is sometimes also made between phylogenetic and evolutionary classification schemes. Phylogenetic methods aim at developing a classification based on an analysis of phylogenetic data, and developing a diagram termed **cladogram** or **phylogenetic tree**, which depicts the genealogical descent of taxa. Biologists practicing this methodology are known as **cladists**, and the field of study as **cladistics**. The term, however, is slowly being replaced by **phylogenetic systematics**. The phylogenetic concepts present a huge diversity of variation, unfortunately often contradictory, leading to different interpretations of similar results. A brief understanding of these is, therefore, necessary before attempting to explore this complex field. Before the development of modern methods of cladistics, the numerical methods were largely used for drawing phylogenetic inferences from the data analysis. The modern Phylogenetic methods, however, integrate the concepts and practices

of numerical taxonomy with cladistic methods. It is, however, essential to understand the concepts of each, and the final integration in phylogeny reconstruction.

PHENETIC METHODS

Numerical taxonomy received a great impetus with the development and advancement of computers. This field of study is also known as **mathematical taxonomy** (Jardine and Sibson, 1971), **taxometrics** (Mayr, 1966), **taximetrics** (Rogers, 1963), **multivariate morphometrics** (Blackith and Reyment, 1971) and **phenetics**. The modern methods of numerical taxonomy had their beginning from the contributions of Sneath (1957), Michener and Sokal (1957), and Sokal and Michener (1958) which culminated in the publication of *Principles of Numerical Taxonomy* (Sokal and Sneath, 1963), with an expanded and updated version *Numerical Taxonomy* (Sneath and Sokal, 1973). The latter authors define Numerical taxonomy as **grouping by numerical methods of taxonomic units into taxa on the basis of their character states**. Before the development of modern methods of cladistics, the numerical methods were also used for drawing phylogenetic inferences from the data analysis.

The last few decades have witnessed a forceful debate on the suitability of the

empirical approach or **operational approach** in systematic studies. Empirical taxonomy forms the classification on the basis of taxonomic judgment based on observation of data and not assumptions. Operational taxonomy, on the other hand, is based on operational methods, experimentation to evaluate the observed data, before a final classification. Numerical taxonomy finds a balance between the two as it is both empirical and operational (Figure 8.1).

It must be remembered that numerical taxonomy does not produce new data or a new system of classification, but is rather a new method of organizing data that could help in better understanding of relationships. Special classifications are based either on one or a few characters or on one set of data. Numerical taxonomy seeks to base classifications on a greater number of characters from many sets of data in an effort to produce an entirely phenetic classification of maximum predictivity.

Principles of Taxometrics

The philosophy of modern methods of numerical taxonomy is based on ideas that were first proposed by the French naturalist Michel Adanson (1763). He rejected the idea of giving more importance to certain characters, and believed that natural taxa are based on the concept of similarity, which is measured by taking all the characters into consideration. The principles of modern numerical taxonomy developed by Sneath and Sokal (1973) are based on the modern interpretation of the Adansonian principles and as such are termed **neo-Adansonian principles**. It would, however, be wrong to visualize Adanson as the founder of numerical taxonomy, because he worked in a different academic environment from that of today, when tools of investigation were much different. These **principles** of numerical taxonomy are enumerated below.

1. The greater the content of information in the taxa of a classification and the more characters it is based upon,

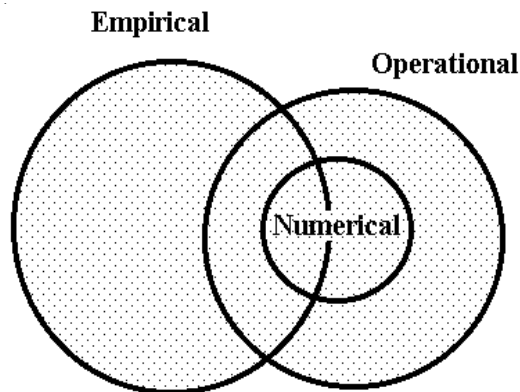


Figure 8.1 Relationship between empirical, operational and numerical taxonomy (after Sneath and Sokal, 1973).

the better a given classification will be.

2. A priori, every character is of equal weight in creating natural taxa.
3. Overall similarity between any two entities is a function of their individual similarities in each of the many characters in which they are being compared.
4. Distinct taxa can be recognized because correlations of characters differ in the groups of organisms under study.
5. Phylogenetic inferences can be made from the taxonomic structures of a group and also from character correlations, given certain assumptions about evolutionary pathways and mechanisms.
6. Taxonomy is viewed and practiced as an empirical science.
7. Classifications are based on phenetic similarity.

The methodology of numerical taxonomy involves the selection of operational units (populations, species, genera, etc., from which the information is collected) and characters. The information from these is recorded, and similarity (and/or distance) between units is determined using various

statistical formulae. The ultimate analysis involves comparison of similarity data and constructing diagrams or models, which provide a summary of the data analysis. These diagrams or models are used for final synthesis and better understanding of the relationships. The **major advantages** of numerical taxonomy over conventional taxonomy include:

1. Numerical taxonomy has the power to integrate data from a variety of sources such as morphology, physiology, phytochemistry, embryology, anatomy, palynology, chromosomes, ultrastructure and micromorphology. This is very difficult to do by conventional taxonomy.
2. Considerable automation of the data processing promotes efficiency and the work can be handled by even less skilled workers.
3. Data coded in numerical form can be integrated with existing data-processing systems in various institutions and used for the creation of descriptions, keys, catalogues, maps and other documents.
4. The methods, being quantitative, provide greater discrimination along the spectrum of taxonomic differences, and can provide better classifications and keys.
5. The creation of explicit data tables for numerical taxonomy necessitates the use of more and better described characters, which will necessarily improve conventional taxonomy as well.
6. The application of numerical taxonomy has posed some fresh questions concerning classification and initiated efforts for re-examination of classification systems.
7. A number of biological and evolutionary concepts have been reinterpreted, thus introducing renewed interest in biological research.

Numerical taxonomy aims at determining **phenetic relationships** between organisms or taxa. Cain and Harrison (1960) de-

finer phenetic relationship as ***an arrangement by overall similarity, based on all available characters without any weighting***. Sneath and Sokal (1973) define **phenetic relationship** as ***similarity (resemblance) based on a set of phenotypic characteristics and not phylogeny of organisms under study***. It is distinct from a cladistic relationship, which is an expression of the recency of common ancestry and is represented by a branching network of ancestor-descendant relationships. Whereas the phenetic relationship is represented by a **phenogram**, the cladistic relationship is depicted through a **cladogram**.

CLADISTIC METHODS

Although phylogenetic diagrams (now appropriately known as **phylograms**) have been used by Bessey (1915), Hutchinson (1959, 1973), and contemporary authors of classification systems to depict the relationships between taxa, the cladograms are distinct in the sense that they are developed using a distinct methodology. This method was first proposed by W. Hennig (1950, 1957), a German zoologist who founded the subject of **phylogenetic systematics**. The term **cladistics** for this methodology was coined by Mayr (1969). An American Botanist, W. H. Wagner, working independently, developed a method of constructing phylogenetic trees, called the groundplan-divergence method, in 1948. Over the years, cladistics has developed into a forceful methodology of developing phylogenetic classifications.

Cladistics is a methodology that attempts to analyse phylogenetic data objectively, in a manner parallel to taxometrics, which analyses phenetic data. Cladistic methods are largely based on the **principle of parsimony** according to which, the most likely evolutionary route is the shortest hypothetical pathway of changes that explains the pattern under observation. Taxa in a truly phylogenetic system should be monophyletic. It has been found that **symplesiomorphy** (possession of primitive or plesiomorphic

character-state in common by two or more taxa) does not necessarily indicate monophyly. **Synapomorphy** (possession of derived or apomorphic character-state in common by two or more taxa), on the other hand, is a more reliable indicative of monophyly. It is thus common to use homologous shared and derived character-states for cladistic studies. Before analysing the methodology of handling data for phylogenetic analysis, it is important to understand the major terms and concepts used in Phylogenetic Systematics.

Phylogenetic Terms

Many important terms have been repeatedly used in discussions on the phylogeny of angiosperms, with diverse interpretation, which has often resulted in different sets of conclusions. A prominent case in point is Melville (1983), who regards the angiosperms as a monophyletic group. His justification—several ancestral forms of the single fossil group *Glossopteridae* gave rise to angiosperms—renders his view as polyphyletic in the eyes of the greater majority of authors who believe in the strict application of the concept of monophyly. The involvement of more than one ancestor makes angiosperms a polyphyletic group, a view that has been firmly rejected. A uniform thorough evaluation of these concepts is necessary for proper understanding of angiosperm phylogeny.

Plesiomorphic and Apomorphic Characters

A central point to the determination of the phylogenetic position of a particular group is the number of primitive (**plesiomorphic**) or advanced (**apomorphic**) characters (although the term character is often used broadly in literature, more appropriately primitive or advanced and similarly plesiomorphic and apomorphic refer to different character-states of a character, and not different characters) that the group contains. In the past, most conclusions on primitiveness were based on circular reasoning:

'These families are primitive because they possess primitive characters (or character-states) and primitive characters (or character-states) are those which are possessed by these primitive families'. Over the recent years, a better understanding of these concepts has become possible. It is generally accepted that evolution has proceeded at different rates in different groups of plants so that among the present-day organisms, some are more advanced than others. The first step in the determination of relative advancement of characters, is to ascertain which characters are plesiomorphic and which are apomorphic. Stebbins (1950) argued that it is wrong to consider the characters as separate entities, since it is through the summation of characters peculiar to an individual, that natural selection operates. Sporne (1974) while agreeing with this, believed that it is scarcely possible initially to avoid thinking in terms of separate characters, which can be treated better statistically. Given insufficient fossil records of the earliest angiosperms, comparative morphology has been largely used to decide the relative advancement of characters. Many doctrines have been proposed but unfortunately most rely on circular reasoning. Some of the important doctrines are described below:

The **Doctrine of conservative regions** holds that certain regions of plants have been less susceptible to environmental influence than others and, therefore, exhibit primitive features. Unfortunately, however, over the years, every part of the plant has been claimed as conservative region. Also, the assumption that a flower is more conservative than the vegetative parts is derived from classifications which are based on this assumption.

The **doctrine of recapitulation** holds that early phases in development are supposed to exhibit primitive features, i.e. '**ontogeny repeats phylogeny**'. Gunderson (1939) used this theory to establish the following evolutionary trends: polypetal to gamopetal (since the petal primordia are initially separate, the tubular portion of the corolla arises later); polysepal to gamosepal;

actinomorphy to zygomorphy and apocarpy to syncarpy. The concept originally applied to animals does not always hold well in plants where ontogeny does not end with embryogeny but continues throughout the adult life. **Neoteny** (persistence of juvenile features in mature organism) is an example wherein a persistent embryonic form represents an advanced condition.

The **doctrine of teratology** was advocated by Sahni (1925), who argued that when a normal equilibrium is upset, an adjustment is often effected by falling back upon the surer basis of past experience. Thus, teratology (abnormality) is seen as reminiscent of some remote ancestor. According to Heslop-Harrison (1952), some teratological phenomena are just likely to be progressive or retrogressive, and each case must be judged on its own merit.

The **doctrine of sequences** advocates that if organisms are arranged in a series in such a way as to show the gradation of a particular organ or structure, then the two ends of the series represent apomorphy and plesiomorphy. The most crucial decision, however, is from which end should the series be read.

The **doctrine of association** advocates that if one structure has evolved from another, then the primitive condition of the derived one will be similar to the general condition of the ancestral structure. Thus, if vessels have evolved from tracheids, then the vessels similar to tracheids (vessels with longer elements, smaller diameter, greater angularity, thinner walls and oblique end walls) represent a more primitive condition than vessels with broader, shorter, more circular elements with horizontal end walls.

The **doctrine of common ground plan** advocates that characters common to all members of a group must have been possessed by the original ancestor and must, therefore, be primitive. The doctrine, however, cannot be applied to angiosperms in which there is an exception for almost every character.

The **doctrine of character correlation** was acknowledged during the second decade of the previous century when it was realized

that certain morphological characters are statistically correlated and the fact can be used in the study of evolution. Sinnott and Bailey (1914) demonstrated a positive correlation between trilacunar node and stipules. Frost (1930) believed that correlation between characters arises because rates of their evolution have been correlated. Sporne (1974) has, however, argued that correlation can be shown to occur even though the rates of evolution of characters are not the same. Within any taxonomic group, primitive characters may be expected to show positive correlation merely because their distribution is not random. By definition, primitive members of that group have retained a relatively high proportion of ancestral (plesiomorphic) characters, while advanced members have dispensed with a relatively high proportion of these same characters—either by loss or replacement with different (apomorphic) characters. It follows, therefore, that the distribution of plesiomorphic characters is displaced towards primitive members, which have a higher proportion of plesiomorphic characters, than the average for the group as a whole. Departure from the random can be statistically calculated in order to establish correlation among characters. Based on these calculations, Sporne (1974) prepared a list of 24 characters in Dicotyledons and 14 in Monocotyledons, which exhibit positive correlation. These characters, because of their distribution, have been categorized as **magnoloid** and **amarylloid**, respectively. Based on the distribution of these characters, Sporne calculated an **advancement index** for each family and projected the placement of different families of angiosperms in the form of a **circular diagram**, with the most primitive families near the centre, and the most advanced along the periphery. That the earliest members of angiosperms are extinct is clear from the fact that none of the present-day families has the advancement index of zero. All living families have advanced in some respects.

The concept of apomorphic and plesiomorphic characters in understanding

the phylogeny of angiosperms has been considerably advanced with the recent development of **cladistic methods**. These employ a distinct methodology, somewhat similar to taxometric methods in certain steps involved, leading to the construction of **cladograms** depicting evolutionary relationships within a group. Certain groups of angiosperms are reported to have a combination of both plesiomorphic and apomorphic characters, a situation known as **heterobathmy**. *Tetracentron* has primitive vesselless wood but the pollen grains are advanced, being tricolpate.

Homology and Analogy

Different organisms resemble one another in certain characters. Taxonomic groups or taxa are constructed based on overall resemblances. The resemblances due to homology are real, whereas those due to analogy are generally superficial. A real understanding of these terms is, thus, necessary in order to keep organisms with superficial resemblance in separate groups. The two terms as such play a very important role in understanding evolutionary biology.

These terms were first used and defined by Owen (1848). He defined **Homology** as *the occurrence of the same organ in different animals under every variety of forms and functions*. He defined **Analogy** as *the occurrence of a part or an organ in one animal which has the same function as another part or organ in a different animal*. If applied to plants, the rhizome of ginger, the corm of colocasia, tuber of potato, and runner of lawn grass are all homologous, as they all represent a stem. The tuber of potato and the tuber of sweet potato, on the other hand, are analogous as the latter represents a root.

Darwin (1959) was the first to apply these terms to both animals and plants. He defined homology as *that relationship between parts which results from their development from corresponding embryonic parts*. The parts of a flower in different plants are thus homologous and these, in turn, are

homologous with leaves because their development is identical.

During the latter half of the present century, phylogenetic interpretation has been applied to these terms. Simpson (1961) defined homology as *the resemblance due to inheritance from a common ancestry*. Analogy, similarly, represents *functional similarity and not due to inheritance from a common ancestry*. Mayr (1969) similarly defined homology as *the occurrence of similar features in two or more organisms, which can be traced to the same feature in the common ancestor of these organisms*. It is, as such, imperative that homology between two organisms can result only from their having evolved from a common ancestor, and the ancestor must also contain the same feature or features for which the two organisms are homologous.

Wiley (1981) has provided a detailed interpretation of these terms. Homology may either be between two characters, two character states, or between two organisms for a

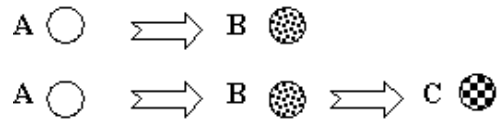


Figure 8.2 Homology between characters (or character states). In the first example, character A is plesiomorphic and B is apomorphic. In the second example, B is apomorphic in relation to A but plesiomorphic in relation to C as all three belong to an evolutionary transformation series.

particular character or character state. **Two characters (or character-states) are homologous if one is directly derived from the other**. Such a series of characters is called an **evolutionary transformation series** (also called **morphoclines** or **phenoclines**). The original, pre-existing character (or character-state) is termed **plesiomorphic** and the derived one as **apomorphic** or **evolutionary novelty**.

Three or more character-states may be homologous if they belong to the same evolutionary transformation series (ovary superior \rightarrow half-inferior \rightarrow inferior). The terms plesiomorphic and apomorphic are, however, relative. In an evolutionary transformation series representing characters A, B and C (Figure 8.2), B is apomorphic in relation to A but it is plesiomorphic in relation to C.

Two or more organisms may be homologous for a particular character (or character-state) if their immediate common ancestor also had this character. Such a character is called **shared homologue**. If the character-state is present in the immediate common ancestor, but not in the earlier ancestor (Figure 8.3), i.e. the character-state is a derived one, the situation is known as **synapomorphy**. If the character-state is present in the immediate common ancestor, as well as in the earlier ancestor, i.e. it is an original character-state, the situation is known as **symplesiomorphy** (note sym-).

The homology between different organisms is termed **special homology**, as represented by different types of leaves in different species of plants. Different leaves in the same plant such as foliage leaves, bracts, floral leaves would also be homologous, representing **serial homology**. The following criteria may be helpful in identifying homology in practice:

1. Morphological similarity with respect to topographic position, geometric position, or position in relation to other parts. A branch, for example, occurs in the axil of a leaf, although it may be modified in different ways.
2. Similar ontogeny.
3. Continuation through intermediates, as for example, the evolution of mammalian ear from gills of fishes, evolution of achene fruit from follicle in Ranunculaceae. Similarly, vessels having evolved from tracheids, the primitive forms of vessels are more like tracheids, with elongated narrower elements with oblique end walls.
4. When the same relatively simple character is found in a large number of species, it is probably homologous in all the species. Sets of characters may similarly be homologous.
5. If two organisms share the characters of sufficient complexity and judged homologous, other characters shared by the organisms are also likely to be homologous.

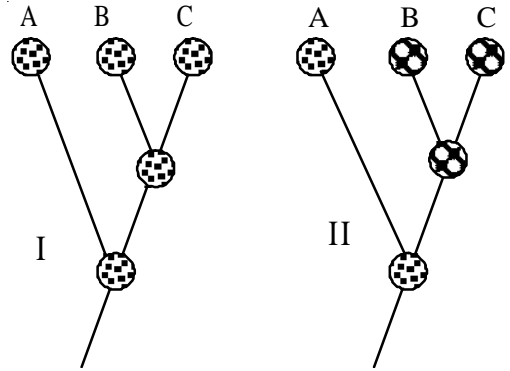


Figure 8.3 Homology between two organisms B and C. In diagram I, similarity is due to symplesiomorphy as the character was unchanged in the previous ancestor. In II, it is due to synapomorphy as the previous ancestor had a plesiomorphic character and the two now share a derived character.

Parallelism and convergence

Unlike homology, if the character shared by two organisms is not traced to a common ancestor, the similarity may be the result of **homoplasy** (sometimes considered synonym of analogy). It can result in three different ways. One, the organisms have a common ancestor but the character-state was not present in their common ancestor (parallelism). It could also result from two different characters in different ancestors evolving into identical character-states (convergence). Similarity could also arise from loss of a particular character (**reversal**), thus reverting to ancestral condition (loss of perianth in some families). All the three situations represent **false synapomorphy**

because the similar character-state is derived and not traced to a common ancestor.

Simpson (1961) defined **parallelism** as the **independent occurrence of similar changes in groups with a common ancestry, and because they had a common ancestry**. The two species *Ranunculus tripartitus* and *R. hederacea* have a similar aquatic habit and dissected leaves and have acquired these characters by parallel evolution. The development of vessels in Gnetales and dicotyledons also represents a case of parallelism.

Convergence implies **increasing similarity between two distinct phyletic lines, either with regard to individual organ or to the whole organism**. The similar features in convergence arise separately in two or more genetically diverse and not closely related taxa or lineages. The similarities have arisen in spite of lack of affinity and have probably been derived from different systems of genes. Examples may be found in the occurrence of pollinia in Asclepiadaceae and Orchidaceae, and the 'switch habit' (circular sheath at nodes) in *Equisetum*, *Ephedra* and *Polygonum*. The concepts of parallelism and convergence are illustrated in Figure 8.4.

Convergence is generally brought about by similar climates and habitats, similar methods of pollination or dispersal. Once the convergence has been identified between two taxa, which have been grouped together, they are separated to make the groups natural and monophyletic. The following criteria may help in the identification of convergence:

1. Convergence commonly results from **adaptation to similar habitats**. Water plants thus usually lack root hairs and root cap but contain air lacunae. Annuals are predominant in deserts, which also have a good number of succulent plants. The gross similarity between certain succulent species of Euphorbiaceae and Cactaceae is a very striking example of convergence.
2. Convergence may also result from **similar modes of pollination** such as

wind pollination in such unrelated families as Poaceae, Salicaceae and Urticaceae, pollinia in Asclepiadaceae and Orchidaceae.

3. Convergence may also be due to **similar modes of dispersal**, as seen in hairy seeds of Asteraceae, Asclepiadaceae and some Malvaceae.
4. Convergence commonly occurs between relatively advanced members of respective groups. *Arenaria* and *Minuartia* form natural groups of species which were earlier placed within the same genus *Arenaria*. The two species *Arenaria leptocladus* and *Minuartia hybrida* show more similarity than between any two species of these two genera. If the similarity is **patristic** (result of common ancestry), then the two species would represent the most primitive members of respective groups (Figure 8.5-I) and it would have been advisable to place all of the species in the same genus *Arenaria*. The

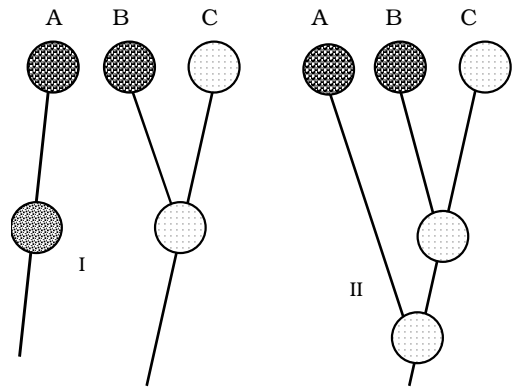


Figure 8.4 Examples of convergence (I) and parallelism (II) between organisms A and B. In convergence, similarity is between organisms derived from different lineages. In parallelism, the ancestor is common but both A and B have evolved an apomorphic character independently. In both cases, similarity represents false synapomorphy. Dissimilarity between B and C in both diagrams is due to divergence.

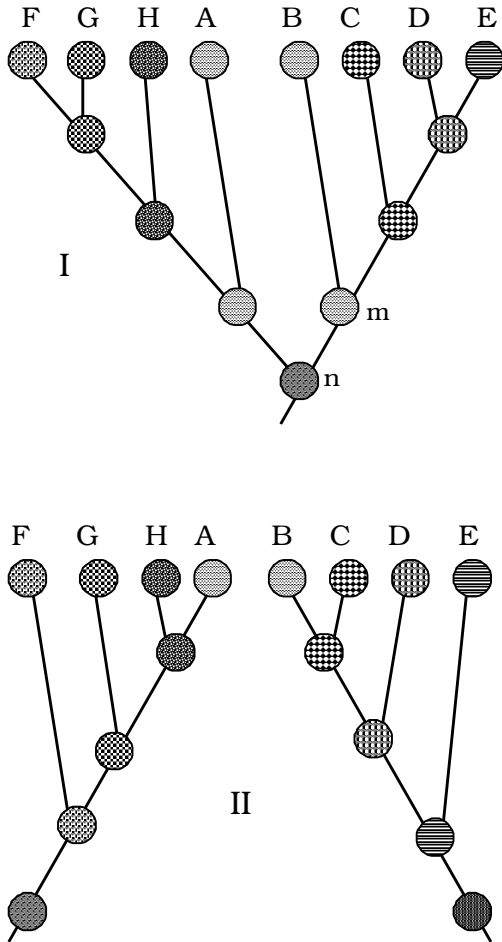


Figure 8.5 Two possible reasons for similarity between species A and B. In (I), A (cf. *Arenaria leptocladus*) and B (cf. *Minuartia hybrida*) are the most primitive members of respective lineages FGHA (cf. *Arenaria*) and BCDE (cf. *Minuartia*). The two lineages have common ancestry and thus constitute a single monophyletic group (cf. *Arenaria* s. l.). In (II), A and B happen to be the most advanced members of the respective groups, the two lineages are distinct and as such similarity between A and B is superficial due to convergence, justifying the independent recognition of two lineages (cf. distinct genera *Arenaria* and *Minuartia*).

studies have shown, however, that these two species are the most specialized in each group (Figure 8.5-II) and thus show convergence. Separation of the two genera is justified, because placing all the species within the same genus *Arenaria* would render the group polyphyletic, a situation that evolutionary biologists avoid.

It is pertinent to mention that although the concepts parallelism and convergence seem to be distinct and theoretically sound, and often easy to apply when discussing **homoplasious** (non-homologous) similarity in the case of closely related organisms (parallelism), or distantly related organisms (convergence), the distinction is not always clear. In Figure 8.5-I, for example if we did not know the evolutionary history of the group before level **m**, there was no way of telling whether all the eight species had a common ancestor or not. For practical reasons, it is always safer to refer homoplasious situations together. Some recent authors like Judd et al., (2002) treat parallelism and convergence as same.

Reversal is a common evolutionary process, wherein loss of a particular character may lead to apparent similarity with ancestral condition. The occurrence of reduced unisexual flowers without perianth or with reduced perianth in Amentiferae was once considered to be primitive situation, but the evidence from wood anatomy, floral anatomy and palynology have shown that apparent simplicity of these flowers is due to evolutionary reduction (reversal), and as such the assumed similarity to angiosperm ancestral condition is representation of homoplasy, a false similarity between an evolutionary advancement (secondary reduction) and ancestral simple condition.

Monophyly, Paraphyly and Polyphyly

These terms have been commonly used in taxonomy and evolutionary literature with such varied interpretation that much confusion has arisen in their application.

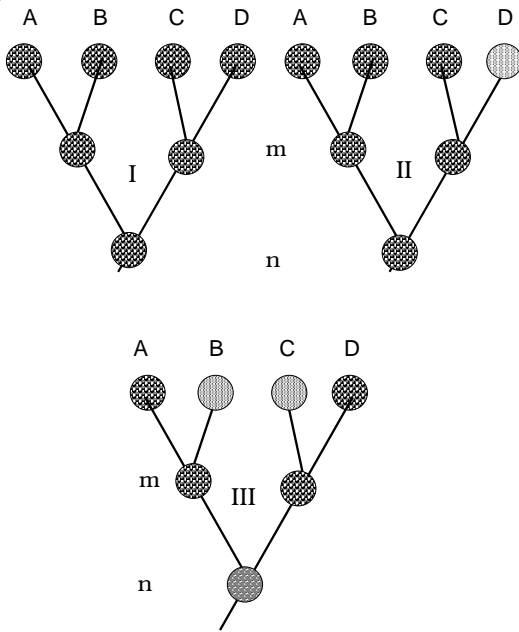


Figure 8.6 Concepts of monophyly, paraphyly and polyphyly. In (I) groups AB and CD are monophyletic as each has a common ancestor at level \underline{m} . Similarly, group ABCD is monophyletic as it has a common ancestor at level \underline{n} . In (II) group ABC is paraphyletic as we are leaving out descendant D of the common ancestor at level \underline{n} . In (III) group BC is polyphyletic as their respective ancestors at level \underline{m} do not belong to this group.

Defined broadly, the terms monophyly (derivation from a single ancestor) and polyphyly (derivation from more than one ancestor) would have different meanings depending upon how far back we are prepared to go in evolutionary history. If life arose only once on Earth, all organisms (even if you place an animal species and a plant species in the same group) are ultimately monophyletic in origin. There is thus a need for a precise definition of these terms, to make them meaningful in taxonomy.

Simpson (1961) defined **monophyly** as *the derivation of a taxon through one or*

more lineages from one immediately ancestral taxon of the same or lower rank. Such a definition would be true if, say, genus B evolved from genus A through one species of the latter, since in that case, the genus B would be monophyletic at the same rank (genus) as well as at the lower (species) rank. On the other hand, if genus B evolved from two species of genus A, it would be monophyletic at the genus level but polyphyletic at the lower rank.

Most authors, however, including Heslop-Harrison (1958) and Hennig (1966), adhere to a stricter interpretation of monophyly, namely the group should have evolved from a single immediately ancestral species which, may be considered as belonging to the group in question. There are thus two different levels of monophyly: a **minimum monophyly** wherein one supraspecific taxon is derived from another of equal rank (Simpson's definition), and a **strict monophyly** wherein one higher taxon is derived from a single evolutionary species.

Mayr (1969) and Melville (1983) follow the concept of minimum monophyly. Most authors, including Heslop-Harrison (1958), Hennig (1966), Ashlock (1971) and Wiley (1981), reject the idea of minimum monophyly. All supraspecific taxa are composed of individual lineages that evolve independent of each other and cannot be ancestral to one another. Only a species can be an ancestor of a taxon. The supraspecific taxa are not biologically meaningful entities and are only evolutionary artifacts.

Hennig (1966) defined a monophyletic group as *a group of species descended from a single ('stem') species, and which includes all the descendants from this species.* Briefly, a monophyletic group comprises all the descendants that at one time belonged to a single species. A useful analysis of Hennig's concept of monophyly was made by Ashlock (1971). He distinguished between two types of monophyletic groups: those that are **holophyletic** when *all descendants of the most recent common ancestor are contained in the group* (monophyletic sensu

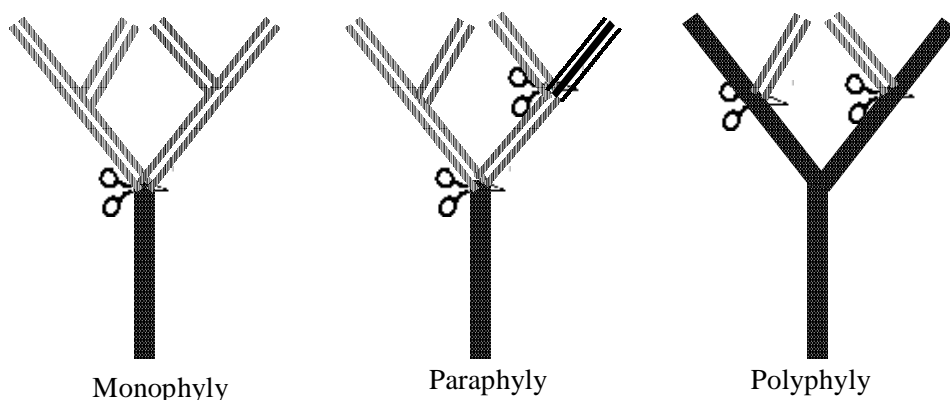


Figure 8.7 The application of cutting rules to distinguish between monophyly, paraphyly and polyphyly. The group is represented by lighter portion of the tree. Monophyletic group can be separated by a single cut below the group, a paraphyletic group by one cut below the group and one or more higher up. A polyphyletic is separated by more than one cut below the group. A monophyletic group represents one complete branch, a paraphyletic group one larger portion of the branch; whereas the polyphyletic group represents more than one pieces of a branch (based on Dahlgren et al., 1985).

Hennig) and those that are **paraphyletic** and **do not contain all descendants of the most recent common ancestor of the group**. A **polyphyletic** group, according to him, is one whose most recent ancestor is not clastically a member of that group. The terms holophyletic and monophyletic are now considered synonymous. Diagrammatic representations of Ashlock's concept of polyphyly, monophyly and paraphyly is presented in Figure 8.6.

An excellent representation of monophyly, paraphyly and polyphyly is presented by '**cutting rules**', devised by Dahlgren and Rasmusen (1983). The distinction is based on how the group is separated from a representative evolutionary tree (Figure 8.7). A **monophyletic group** is **separated by a single cut below the group**, i.e. it represents **one complete branch**. A **paraphyletic group** is **separated by one cut below the group and one or more cuts higher up**, i.e. it represents **one piece of a branch**. A **polyphyletic group**, on the other hand, is **separated by more than one cut below the**

group, i.e., it represents **more than one piece of a branch**.

Gerhard Haszprunar (1987) introduced the term orthophyletic while discussing the phylogeny of Gastropods. An **orthophyletic** group is a stem group, i.e. a group that is paraphyletic because a single clade (the crown group), has been excluded. The term has not been followed in other groups, especially in botanical systematics. Sosef (1997) compares the existent hierarchical models of classification. He argues that a phylogenetic tree can be subdivided according to a monophyletic hierarchical model, in which only monophyletic units figure or, according to a 'Linnaean' hierarchical model, in which both mono- and paraphyletic units occur. Most present-day phylogeneticists try to fit the monophyletic model within the set of nomenclatural conventions that fit the Linnaean model. However, the two models are intrinsically incongruent. The monophyletic model requires a system of classification of its own, at variance with currently accepted conventions. Since, however, the mono-

phyletic model is unable to cope with reticulate evolutionary relationships; it is unsuited for the classification of nature. The Linnaean model is to be preferred. This renders the acceptance of paraphyletic supraspecific taxa inevitable.

As is true for the distinction between parallelism and convergence, similarly, the concepts of paraphyly and polyphyly (both of which are rejected by modern phylogenetic systematics while constructing classification), hold good, when the former is applied to a group of closely related organisms and latter to distantly related organisms. The concepts become ambiguous when a small group of organisms is considered. In Figure 8.6-III, taxa B and C—if brought together—would form

a polyphyletic group, because they are derived from two separate ancestors at level **m**. If, however, A, B, and C are under one group, B and C would still now be components of a paraphyletic group, because one descendant of the common ancestor at level **n** is kept out of the group. A natural group would be one, which includes all descendents of the common ancestor, or the group is monophyletic.

Phylogenetic Diagrams

The affinities between the various groups of plants are commonly depicted with the help of diagrams, with several innovations. These diagrams also help in understanding the classification of included taxa. An

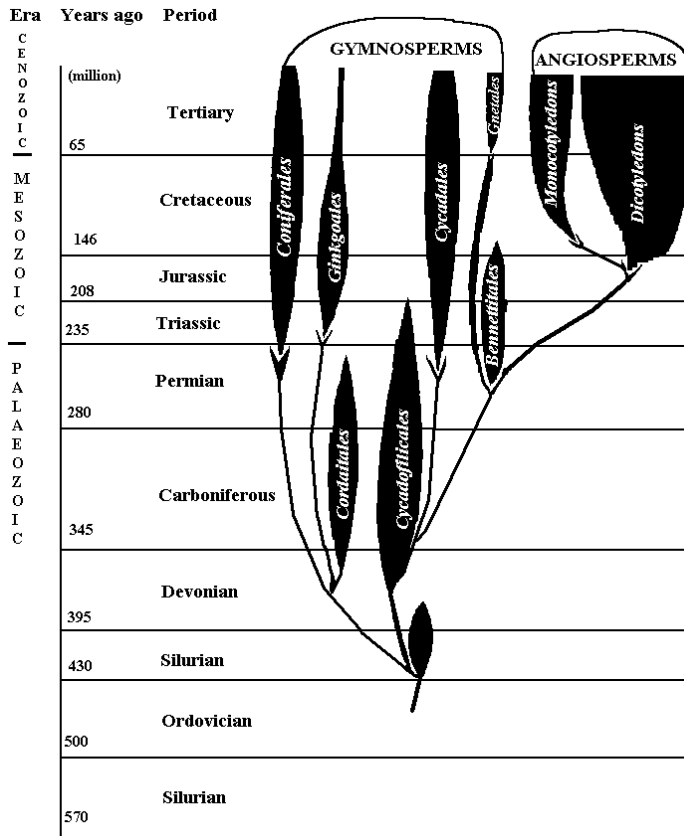


Figure 8.8 A phylogenetic tree representing the evolutionary history of plants including angiosperms. The vertical axis represents the geological time scale. Only extant (living) plants are shown reaching the top.

understanding of these terms is necessary for a correct interpretation of putative relationships. These branching diagrams are broadly known as **dendrograms**. Any diagram showing the evolutionary history of a group in the form of branches arising from one or more points has often been referred to as a **phylogenetic tree**, but the use of terms is now becoming more precise, and more innovative diagrams are being developed often providing useful information about different taxa mapped in the diagram.

The most common form of diagram is one where the length of branch indicates the degree of apomorphy. Such diagrams were sometimes classified as **cladograms** (Stace, 1980), but the term has now been restricted to diagrams constructed through the distinct methodology of cladistics (Stace, 1989). Diagrams with vertical axis representing the degree of apomorphy are now more appropriately known as **phylograms**. The earliest well-known example of such a phylogram is '**Bessey's cactus**' (see Fig 10.11). In such diagrams the most primitive groups end near the base and the most advanced reach the farthest distance.

Hutchinson (1959, 1973) presented his phylogram in the form of a line diagram (figure 10.13). The recent classifications of Takhtajan (1966, 1980, 1987) and Cronquist (1981, 1988) have more innovative phylograms in which the groups are depicted in the form of balloons or bubbles whose size corresponds to the number of species in the group (an approach also found in Besseyan cactus). Such phylograms thus not only depict phylogenetic relationships between the groups, they also show the degree of advancement as also the relative number of species in different groups. Such diagrams have been popularly known as **bubble diagrams**. The bubble diagram of Takhtajan (Figure 10.16) is more detailed and shows the relationship of the orders within the 'bubble'; as mentioned earlier, Woodland (1991) aptly described it as '**Takhtajan's flower garden**'.

The **phylogenetic tree** is a commonly used diagram in relating the phylogenetic

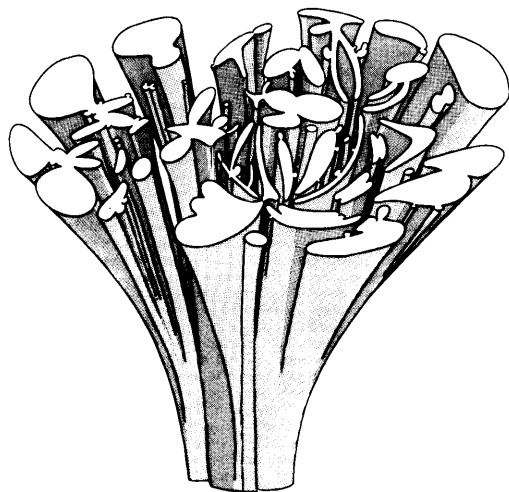


Figure 8.9 Phylogenetic tree of angiosperms presented by Dahlgren (1975) with a section of the top (subsequently named phylogenetic 'shrub' by Dahlgren, 1977).

history. The vertical axis in such a diagram represents the geological time scale. In such a diagram, the origin of a group is depicted by the branch diverging from the main stock and its disappearance by the branch termination. Branches representing the fossil groups end in the geological time when the group became extinct, whereas the extant plant groups extend up to the top of the tree. As already mentioned, the relative advancement of the living groups is indicated by their distance from the centre, primitive groups being near the centre, and advanced groups towards the periphery. A phylogenetic tree representing possible relationships and the evolutionary history of seed plants is presented in Figure 8.8.

Dahlgren (1975) presented the phylogenetic tree (preferred to call it phylogenetic 'shrub' in 1977) of flowering plants with all extant groups reaching the top, and the cross-section of the top of the phylogenetic tree was shown as top plane of this diagram (Figure 8.9). In subsequent schemes of Dahlgren (1977, 1983, 1989), the branching portion of the diagram was dropped and only

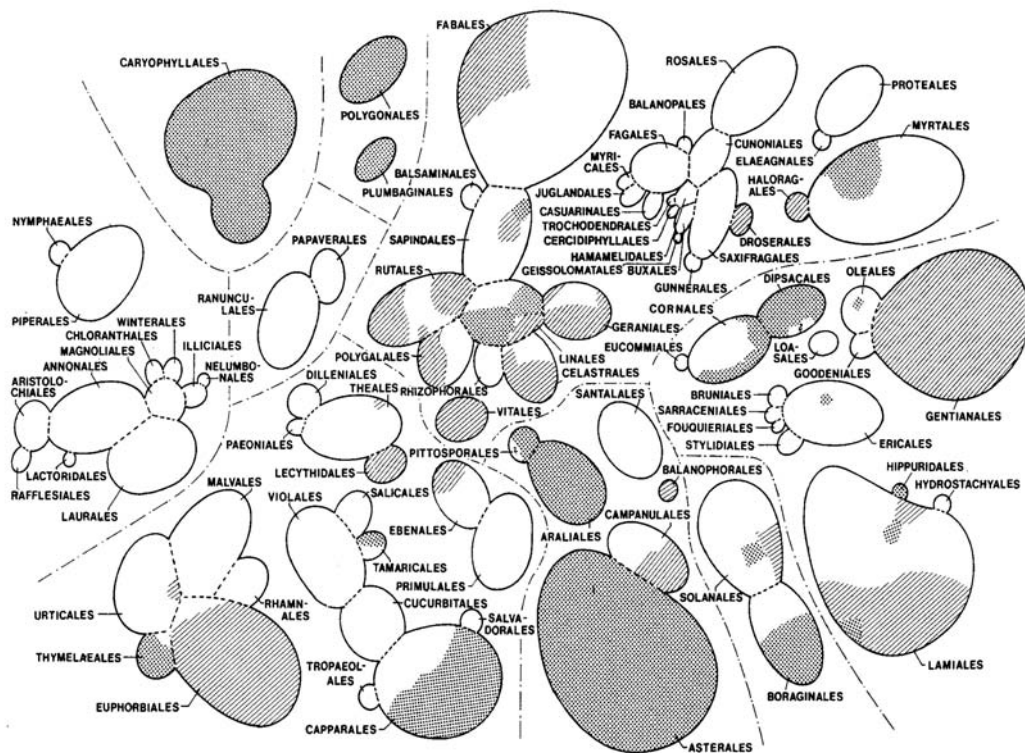


Figure 8.10 Mapping of pollen grain dispersal stage in different dicotyledons on a two-dimensional diagram (Dahlgrenogram) of Dahlgren, representing transverse section through the top of a phylogenetic shrub. Pollen grain dispersal in 2-celled stage (unshaded), 3-celled stage (dotted), or mixed (hatched). (Courtesy Gertrud Dahlgren).

the top plane (cross-section of the top) presented as a two-dimensional diagram (Figure 8.10), and this has been very useful in mapping the distribution of various characters in different groups of angiosperms, and the comparison of these provides a good measure of correspondence of various characters in phylogeny. This diagram has been popularly known as ‘**Dahlgrenogram**’.

Thorne’s diagram (2000) is similarly the top view of a **phylogenetic shrub** (Figure 10.23), in which the centre representing the extinct primitive angiosperms, now absent, is empty.

A **cladogram** represents an evolutionary diagram utilizing cladistic methodology, which attempts to find the shortest hypo-

thetical pathway of changes within a group that explains the present phenetic pattern, using the **principle of parsimony**. A cladogram is a representation of the inferred historical connections between the entities as evidenced by synapomorphies. The vertical axis of the cladogram is always an implied, but usually non-absolute time scale. Cladograms are ancestor-descendant sequences of populations. Each bifurcation of the cladogram represents a past speciation that resulted in two separate lineages.

It must be pointed out, however, that considerable confusion still exists between application of the terms cladogram and phylogenetic tree. Wiley (1981) defines a cladogram as **a branching diagram of entities**

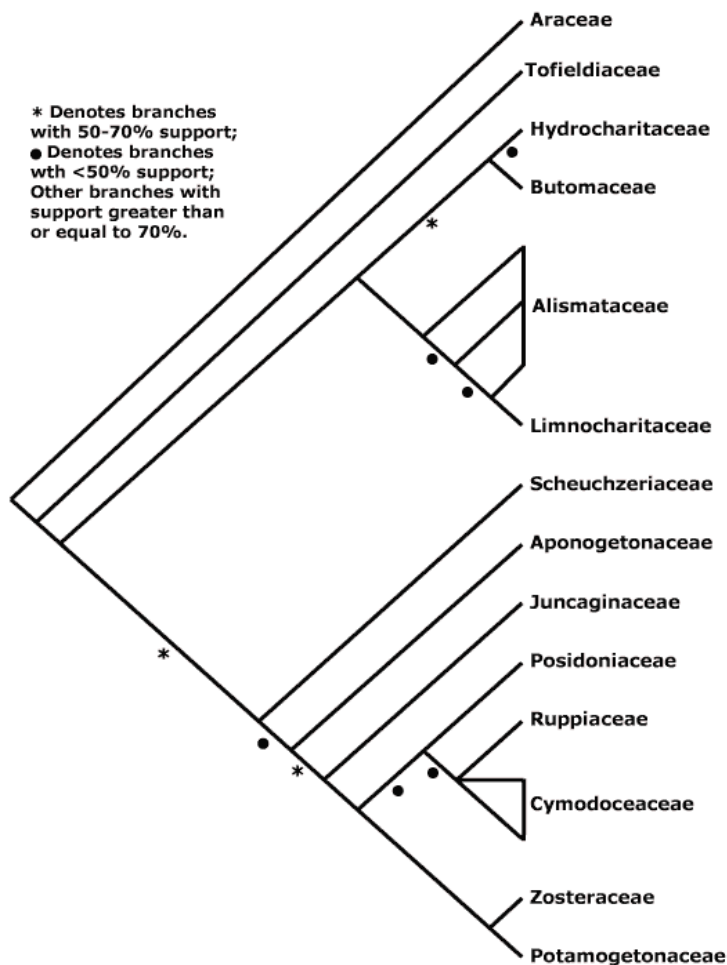


Figure 8.11 Tree (cladogram) for different families of the order Alismatales. Support indicated for branches refers to bootstrap support, discussed in subsequent pages. (Reproduced from APweb version 7 (June, 2008), with permission from Dr P. F. Stevens.)

where the branching is based on inferred historical connections between the entities as evidenced by synapomorphies. It is, thus, a phylogenetic or historical dendrogram. He defines a phylogenetic tree as **a branching diagram portraying hypothesized genealogical ties and sequences of historical events linking individual organisms, populations, or taxa.** At the species and population level, the number of possible phylogenetic trees could be more than cladograms for particular character

changes, depending on which species is ancestral and being relegated lower down on the vertical axis. In the case of higher taxa, the number of cladograms and phylogenetic trees could possibly be equal, because higher taxa cannot be ancestral to other higher taxa since they are not units of evolution but historical units composed of separately evolving species.

Over the recent years, it has been thus becoming increasingly common to construct evolutionary diagrams using cladistic

methodology, assuming that these character-state changes (represented as evolutionary scale or tree length) correspond to the geological time scale, and call these evolutionary diagrams as **evolutionary tree** (Judd et al., 2008), **phylogenetic tree**, or simply **tree** (Stevens, 2008), synonymous with a **cladogram**.

A **Phenogram** is a diagram constructed on the basis of numerical analysis of phenetic data. Such a diagram is the result of utilization of a large number of characters, usually from all available fields, and involves calculating the similarity between taxa and constructing a diagram through **cluster analysis**. Such a diagram (Figure 8.20) is very useful, firstly because it is based on a large number of characters, and secondly because a hierarchical classification can be achieved by deciding upon the threshold levels of similarity between taxa assigned to various ranks.

It must be pointed out that the modern phylogenetic methods, which aim at constructing phylogenetic trees, also sometimes use large number of characters for comparison, especially when dealing with morphological data, and there seem to be a lot of similarities in data handling and computation, but are unique in the utilization of evolutionary markers and, consequently, produce slightly different results. With the incorporation of distance methods in the construction of trees, the classical difference between the terms is largely disappearing. Modern cladistic programs develop trees in which branch lengths are indicated, and plotting programs offer the choice to indicate branch lengths (and often called phylograms) or not. In latter case branches may be square (line running vertically and horizontally- and often called phenogram; Figure 8:21) or V-shaped (cladogram; Figure 8.11). These may be presented as upright or as horizontal trees (**prostrate trees**). Modern trees contain information about evolutionary markers such as bootstrap support, branch length, and Bremer support, as discussed in subsequent pages.

Phylogeny and Classification

The construction of phylogenetic classification involves two distinct steps: determining the **phylogeny** or evolutionary history of a group, and construction classification on the basis of this history. Imagine a **lineage** (or **clade**- a group of individuals producing successively, similar and genetically related individuals, generally represented by lines in a cladogram) with woody habit, alternate leaves, cymose inflorescence, 5 red petals, 5 stamens, 2 free carpels, and dry fruit with many seeds. Over a period of time, some population acquires herbaceous habit and the original lineage splits into two, one with woody habit and the second with herbaceous habit (Figure 8.12). In the lineage with woody habit, one lineage emerges with fused carpels, while the other loses one of the two carpels. The one with fused carpels loses 3 of the five stamens in one or more populations, and that with a single carpel doubles the number of stamens to ten in one or more populations. The herbaceous lineage, similarly, splits into one with yellow petals and one with white petals, the former developing fleshy fruits in one or more populations, and the latter having the number of seeds reduced to one in some populations. The present descendants of the original ancestor are thus represented by eight lineages, which have developed a few apomorphic character-states, but also share plesiomorphic character-states such as alternate leaves, 5 petals, and cymose inflorescence. There must be hundreds of more plesiomorphic states, but of little significance in classification, as the above three. Note that the ancestral species at level I, II (woody habit, 2 free carpels), IA (herbaceous habit, red petals), III (herbaceous habit, yellow petals and free carpels and dry fruit), and IV (herbaceous habit, white petals, 5 stamens and many seeds) and have disappeared, whereas those at level V, and VI are still represented (although with minor changes) in the form of E and G, respectively. Also note that united carpels have arisen twice independently. The same is also true for the loss of three stamens.

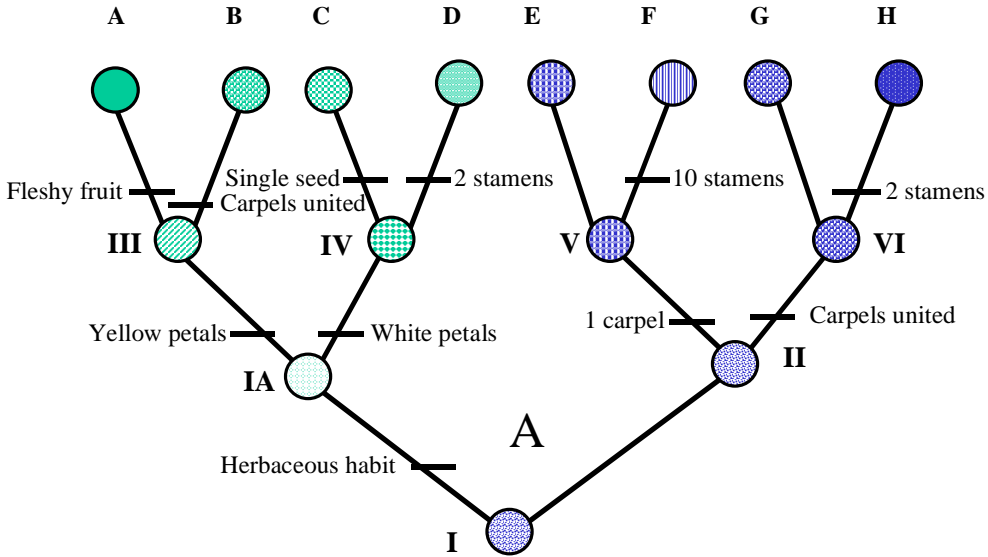


Figure 8.12 Evolutionary history of a hypothetical group of organisms which started with the ancestral species with woody habit, alternate leaves, cymose inflorescence, 5 red petals, 5 stamens, 2 free carpels and dry fruit with many seeds. Eleven character state transformations at different stages have resulted in 8 present day species. Note that two of the changes (carpel union and loss of three stamens) have occurred twice, and as such only nine **genetic switches** are involved. The ancestral species at levels I, IA, II, III and IV have disappeared.

Having known the evolutionary history of the group, we could use synapomorphy and the concept of monophyly. Assuming that all eight lineages (groups of populations) are sufficiently distinct to be recognized as distinct species, we would have eight species. The simplest way would be to group these eight species into four genera, each having a common ancestor. Two of these common ancestors have disappeared, but two are still living and would also be included in the respective genera (it will be more appropriate to regard E as ancestral to F and G ancestral to H). These could be further assembled into two families of four species each (two genera each) having a common ancestor at level IA and II, and these two families into one order with a common ancestor at level I. Please note that common ancestor at level I, IA, II, III and IV are also no longer living.

The second option would be to include ABCD in one genus, and EFGH in another

genus, and include all 8 species (2 genera) in one family (and, of course, depending upon the degree of diversity from related families, this could still be a constituent of a monotypic order). The third option would be to have a single genus of eight species.

Note the importance of synapomorphy in determining monophyly. Character-states alternate leaves, cymose inflorescence and 5 petals (character-states of different characters not same) have been passed unchanged in all the eight descendants (species), which as such are **symplesiomorphic** for each of these, and this symplesiomorphy will be valid between any two (or more) species that you choose to combine into a genus, say, D and E, or C and F, or say ABCDE. On the other hand, if we consider only synapomorphy, the monophyly is easily deciphered. A and B are accordingly synapomorphic for yellow petals, C and D for white petals, E and F for one carpel and

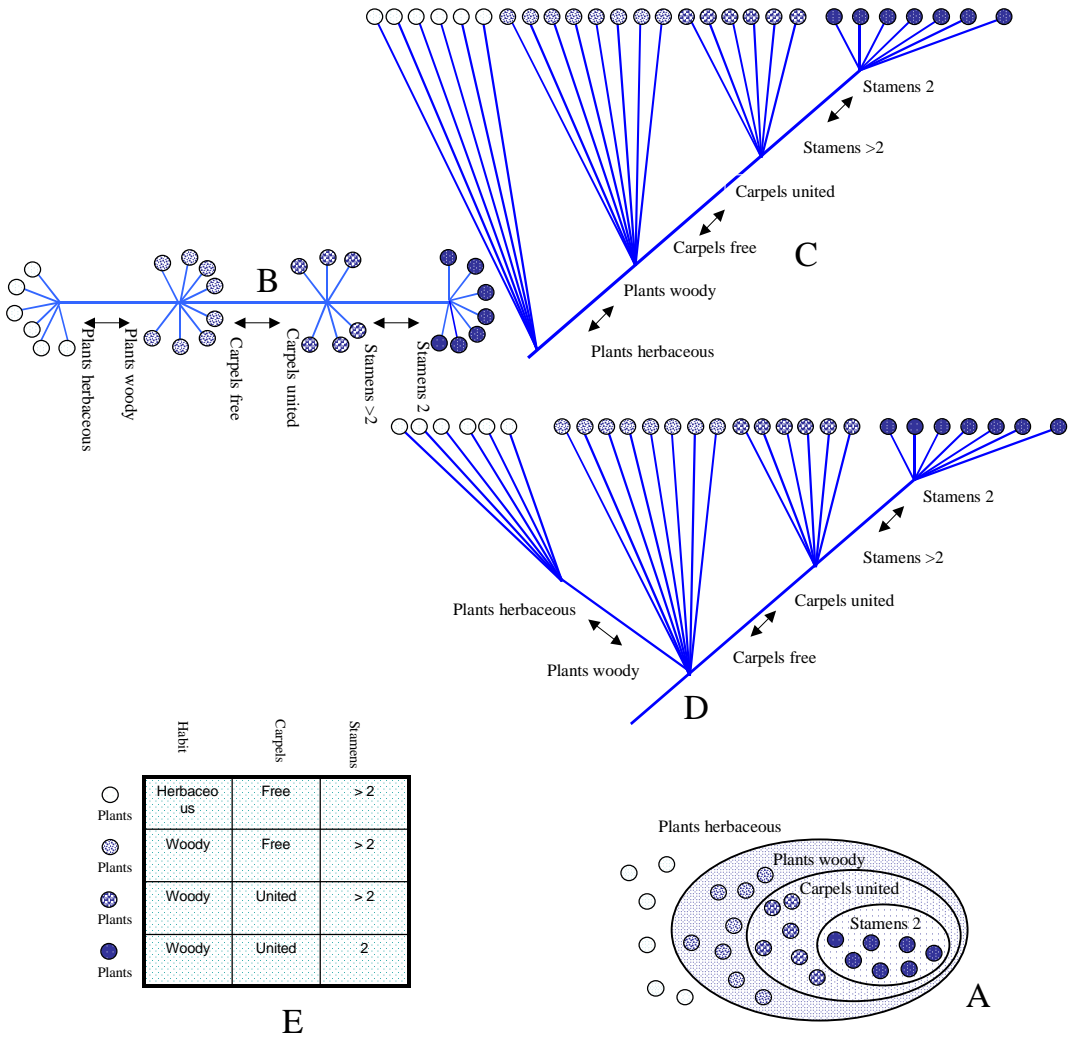


Figure 8.13 Diagrams based on evolutionary pattern depicted in Figure 10.11. **A:** Venn diagram based on the assumption that there are 21 woody species of which, 13 are with united carpels, and 8 with free carpels. Of the 13 with united carpels, 7 have 2 stamens whereas 6 have more stamens. **B:** An unrooted tree based on Venn diagram. **C:** A possible rooted tree, if evolutionary history of the group was not known, 15 possible rooted trees could be drawn. **D:** Rooted tree based on knowledge that herbaceous habit arose from woody habit, and we know further evolution of woody species. Other possibilities are discussed subsequently. **E:** Data matrix of above, where the number of species are not indicated.

G and H for united carpels. Similarly, A, B, C and D are synapomorphic for herbaceous habit. The development of fleshy fruit in A, single seed in C, 10 stamens in F represent the occurrence of a derived character state in a single taxon, and termed as **autapomorphy**. Such character states are not helpful in cladogram construction, although indicative of divergence. Development of 2 stamens in D and H independently represents **homoplasy**, and may lead to artificial grouping of these two, if history of the group was not properly known. It must, however, be noted that symplesiomorphy may sometimes be helpful in detecting monophyly, especially where in some other taxa, it has evolved into another character-state. As such, out of the four plesiomorphic character-states listed here, only the woody habit has changed to herbaceous habit, and as such in the remaining taxa (E, F, G and H) symplesiomorphy of woody character-state identifies monophyly of the group ABCD. It must be remembered that synapomorphy and symplesiomorphy are a reflection of **homology** for a particular character-state (or more than one character-states, each belonging to a different character).

All above options would render (genus, family or order) monophyletic groups, the ultimate goal of phylogenetic systematics. Any other options won't work. Keeping D and E in one genus (or CDE or DEF) would make it **polyphyletic**, promptly rejected once detected, because the group is derived from more than one ancestor. Keeping ABC under one genus, or FGH under one genus would make **paraphyletic** groups because we are not including all the descendents of the common ancestor (we are leaving out D in first genus and E in second). In the same way, putting more than four species (but less than eight) under the same genus would make it paraphyletic. Paraphyletic taxa are strongly opposed by phylogenetic systematists; the classical case is the demise of traditional division of angiosperms into monocots and dicots, over the last decade.

It must be noted that all eight species—whatever way you classify—share alternate

leaves, cymose inflorescence and five petals. The species at level IA and above additionally share woody habit, VI and above additionally united carpels, V and above one carpel (and not fused carpels). Similarly, species at level II and above share herbaceous habit (instead of woody habit) in addition to three common, at level III and above additionally yellow petals, and at level IV and above additionally (in place of yellow petals) white petals.

The situation depicted above can be more easily represented through the concept of nested groups, more conveniently represented as a set of ovals, a **Venn diagram** (Figure 8.13A). The diagram drawn here is based on the assumption that we have information from a large group in which there are 21 woody species of which 13 are with united carpels and 8 with free carpels. Of the 13 with united carpels 7 have 2 stamens where as 6 have more stamens.

The information is presented in the form of an **unrooted network (unrooted tree)** (Figure 8.13B). The herbaceous species are shown towards the left of left double arrow, and the woody species towards the right. Similarly, the species towards the left of the middle double arrow are with free carpels while those towards the right with fused carpels. The species towards the left of the right double arrow have more than 2 stamens and those towards the right just 2 stamens.

It must be noted that in constructing the above Venn diagram and the unrooted network, only three character-state transformations are accounted for. We have completely left out grouping of herbaceous plants and the woody plants with a single carpel. Inclusion of these would make the diagrams much more complicated, and present several alternatives. Also, the more meaningful trees have to be **rooted** (the most primitive end at the base), to reflect the phylogeny. Even with the phylogenetic history of the group known, there could be several variations of the rooted tree, two simple ones being shown in the Figure 8.13C and 8.13D. If we did not know the evolutionary history of the group, a number of variations would be

possible, depending upon which character-state is plesiomorphic, and which character (habit, carpel fusion or stamen number) forms the root, and what would be the sequence of the character changes on the tree.

The character-states chosen for analysis should necessarily be homologous (one derived from another) and non-overlapping. The analysis becomes more meaningful when it is established that the evolution of a particular character-state has been the result of a corresponding **genetic change**, and not a mere plastic environmental influence. This fact underlies the importance of the emerging field of molecular analysis in **phylogenetic systematics**. It is believed that the recognition of molecular character-states (nucleotide sequences) is often easier and more precise, although there are always accompanying problems.

The problem with vascular plants, especially the angiosperms, is that we know very little about their evolutionary history. The fossil records, which generally give fair information about evolution, are very scarcely represented. What we have available with us is a mixture of primitive, moderately advanced and advanced groups. Almost each group has some plesiomorphic and some apomorphic character-states, and relative proportion of one or the another delimit the relative advancement of various groups. Attempt to reconstruct the evolutionary history of the group involves comparative study its living members, sorting out plesiomorphic and apomorphic character states, and distribution of these in various members. Once the evolutionary history of the group has been constructed, monophyletic groups at various levels of inclusiveness are identified, assigned ranks, and given appropriate names, to arrive at a working system of classification.

PHYLOGENETIC DATA ANALYSIS

The methodology of cladistics with incorporation of numerical methods involves a number of steps.

Taxa-Operational Units

The first step in data analysis involves the selection of Taxa for data collection, often called Operational Taxonomic Units (OTUs) in Taxometrics, Operational Evolutionary Units (OEU) in cladistics, referring to the sample from which the data is collected. Although it would be ideal to select different individuals of a population, practical considerations make it necessary to select the members of the next lower rank. Thus, for the analysis of a species would need selection of various populations, for the study of a genus they would be different species, and for a family they would be different genera. It is not advisable, however, to use genera and higher ranks, as the majority of characters would show variation from one species to another and thus would not be suitable for comparison. The practical solution would be to use one representative of each taxon. Thus, if a family is to be analysed and its genera to be compared, the data from one representative species of each genus can be used for analysis. Once the taxa are selected, a list of such taxa is prepared. A unique feature of cladistic studies, however, is that the list of taxa generally includes a **hypothetical ancestor**, the comparison with which reveals crucial phylogenetic information, and is used for rooting of the tree. It is, increasingly being realized that only a species is the valid evolutionary entity, and all taxa at higher ranks are artifacts, constructed for the sake of convenience. A meaningful analysis would always be one derived from data from various species (taken from populations) and not any higher rank directly.

Characters

A conventional definition of a taxonomic character is **a characteristic that distinguishes one taxon from another**. Thus, white flowers may distinguish one species from another with red flowers. Hence, the white flower is one character and red flower another. A more practical definition espoused

by numerical taxonomists defines character (Michener and Sokal, 1957) as **a feature, which varies from one organism to another**. By this second definition, flower colour (and not white flower or red flower) is a **character**, and the white flower and red flower are its two **character-states**. Some authors (Colless, 1967) use the term **attribute** for character-state but the two are not always synonymous. When selecting a character for numerical analysis, it is important to select a **unit character**, which may be defined as **a taxonomic character of two or more states, which within the study at hand cannot be subdivided logically, except for the subdivision brought about by the method of coding**. Thus, trichome type may be glandular or eglandular. A glandular trichome may be sessile or stalked. An eglandular trichome may, similarly, be unbranched or branched. In such a case, a glandular trichome may be recognized as a unit character and an eglandular trichome as another unit character. On the other hand, if all glandular trichomes in OTUs are of the same type and all eglandular trichomes are of the same type, the trichome type may be selected as a unit character.

The first step in the handling of characters is to make a list of unit characters. A preliminary step involves **character compatibility** study in which each character is examined to determine the proper sequence of character-state changes that take place as the evolution progresses (**morphoclines** or **transformation series**). The list should include all such characters concerning which information is available. **A priori**, all characters should be weighted equally (no weighting to be given to characters). Although some authors advocate that some characters should subsequently be assigned more weightage than others (**a posteriori weighting**), such considerations generally get nullified when a large number of characters is used. It is generally opined that numerical studies should involve not less than 60 characters, but more than 80 are desirable. For practical consideration, there may be some characters concerning which

information is not available (a large number of plants in a population are not in fruit) or the information is irrelevant (trichome type if a large number of plants are without trichomes), or the characters which show a much greater variation within the same taxon. Such characters are omitted from the list. This constitutes **residual weighting** of characters. The characters (leaves, bracts, carpels) or character states (simple leaf, palmate compound leaf, pinnate compound leaf) chosen should also be homologous, in terms of sharing common ancestry or belonging to same evolutionary transformation series. The 'petals' of *Anemone* are modified sepals and thus not homologous with the petals of *Ranunculus* and hence not comparable. Similarly, the tuber of sweet potato (a modified root) cannot be compared with the tuber of potato (a modified stem).

Binary and multistate characters

The characters most suitable for computer handling are **two-state (binary or presence-absence)** characters (habit woody or herbaceous). However, all characters may not be two-state. They may be qualitative multistate (flowers white, red, blue) or quantitative multistate (leaves two, three, four, five at each node). Such multistate characters can be converted into two-state (flowers white or coloured; leaves four or more vs leaves less than four). Or else the characters may be split (flowers white vs not white, red vs not red, blue vs not blue; leaves two vs not two, three vs not three and so on). Such a splitting may, however, give more weightage to one original character (flower colour or number of leaves). It is essential that different character states identified are **discrete** or discontinuous from one another. Discreteness of character states can be evaluated by comparing the means, ranges, and standard deviations of each character for all taxa in analysis. Additionally t-tests and multivariate analysis may also be used for evaluating character state discreteness.

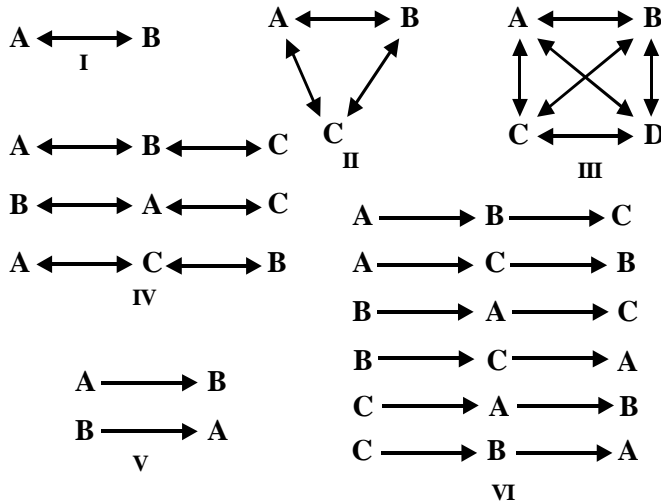


Figure 8.14 Ordering and polarity of character states. I: Binary character with single possible switch. II: Unordered three-state character with single possible switch. III: Unordered Four-state character with single possible switch. IV: Ordered three-state character with two possible switches and three possible morphoclines. V: Polarized binary character with two possible morphoclines. VI: Ordered and Polarized three-state character with 6 possible morphoclines.

Ordering of Character-states

A binary character will have single step or **switch** (Figure 8.14-I) necessary for change. The minimum number of switches possible (**Wagner parsimony**) in a multistate character will depend whether the character states are **ordered** or left **unordered**. In an unordered transformation series each character state can evolve into every other character state, and represents a single switch (Figure 8.14-II, III). A three-state character will have two switches or steps, and three possible morphoclines (Figure 8.14-IV), four-state character three switches and several morphoclines. Whereas ordering of two-state characters is relatively easy, multi-state characters are often difficult to order, and changes may often be reversible, and it is advisable to leave them **unordered**, and identify only one switch (**Fitch parsimony**). The molecular characters are different DNA sequences, that may differ in having one of the four bases (adenine, thymine, guanine

and cytosine) at a particular locus, and as such present four character-states. As reversals are common in these, these are always left unordered.

Assigning Polarity

It is, however, necessary to determine the relative ancestry of the character-states, or the assignment of **polarity**. The designation of polarity is often one of the more difficult and uncertain aspects of phylogenetic analysis. For this, the comparison may be made within the concerned group (**in-group comparison**) or relatives outside the group (**out-group comparison**). The latter may often provide useful information, especially when the out-group used is the **sister-group** of the concerned group. If two character-states of a character are found in a single monophyletic group, the state that is also found in a sister-group is likely to be plesiomorphic and that found only within the concerned monophyletic group is likely to be apomorphic.

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

I

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

II

	A	B	C	D
A	0	1	1	1
B	1	0	1	1
C	1	1	0	1
D	1	1	1	0

III

	A	B
A	0	1
B	1	0

IV

	A	B	C	D
A	0	1	5	5
B	1	0	5	5
C	5	5	0	1
D	5	5	1	0

V

Figure 8.15 Data matrix of coded character states. I: Ordered three-state character. II: ordered four-state character. III: Unordered character. IV: Binary character. V: Differential weighting to character state changes; imagine A and B represent Purines (Adenine and Guanine), C and D Pyrimidines (Cytosine and Thymine), purine to purine or pyrimidine to pyrimidine change (transition) is given 1 step weight, but purine to pyrimidine change or reverse (transversion) given 5 steps weight.

Ingroup comparison (also known as **common ground plan** or **commonality principle**) is based on the presumption that in a given group (presumably monophyletic), the primitive structure would tend to be more common. Thus all 8 species of cladogram in Figure 8.12 share plesiomorphic character states: alternate leaves, cymose inflorescence and five petals. Five species have plesiomorphic 5 stamens, two derived 2 stamens and one with 10 derived stamens. Similarly four species have red petals' and two each with white and yellow petals. It is assumed that the evolution of a derived condition will occur in only one of potentially numerous lineages of the group; thus the ancestral condition will tend to be in the majority. As is evident from Figure 8.14, the number of possible morphoclines increases

after the polarity criterion is included and the selection of single appropriate morphocline representing the true sequence even more challenging.

Character Weighting and Coding

The **coding** of character states is done by assigning non-negative integer values. Binary characters are conveniently assigned 0 and 1 for two states. If possible to distinguish, plesiomorphic state is assigned 0 and apomorphic state 1 code (Figure 8.15-IV). It is often assumed that whereas the same character-state may arise more than once within a group between closely related species (**parallelism**), or between remotely related species (**convergence**; the distinction

Table 8.1 A portion of the data matrix with hypothetical **t** OTUs and **n** characters. Binary coding involves for state a and 1 for state b. The NC code stands for characters not comparable for that OTU. In this analysis a total of 100 characters were used but only nine are pictured here.

OTUs (<i>t</i>)	Characters →								
	Habit 0-woody 1-herbaceous	Fruit 0-follicle 1-achene	Ovary 0-superior 1-inferior	Leaves 0-simple 1-compound	Habitat 0-terrestrial 1-aquatic	Pollen 1-triporate 0-monosulcate	Ovule 1-unitegmic 0-bitegmic	Carpels 0-free 1-united	Plastids 1-PI-type 0-PII-type
1.	1	0	1	1	0	1	1	1	1
2.	1	0	1	1	1	1	0	0	1
3.	0	NC	0	1	0	0	1	1	1
4.	1	1	1	0	1	0	0	0	0
5.	1	0	1	1	0	1	1	0	1
6.	1	1	0	1	1	NC	0	1	0
7.	0	1	1	0	1	1	0	0	1
8.	0	0	0	1	0	0	1	1	1
9.	1	1	1	0	0	1	1	0	0
10.	0	0	0	1	1	0	1	1	1
11.	1	0	1	1	0	1	0	0	1
12.	1	1	0	0	1	1	0	0	1
13.	0	0	1	0	1	0	1	0	1
14.	1	0	1	0	1	0	1	0	1
15.	0	1	1	0	0	1	1	0	0

between parallelism and convergence is sometimes omitted) for a simple character, it is highly unlikely for more complex characters. It is also assumed that whereas many genes must change in order to create a morphological structure, one gene change is enough for its loss (**reversal**). This **Dollo's law** is taken into account when choosing trees, gains of structures counted more than losses, a process known as **Dollo parsimony**. Such **weighting** of characters is often common in phylogenetic analysis. In transformation series leaf simple → pinnately lobed → pinnately compound, the development of pinnate compound leaf from simple leaf occurred in two steps, and needs to be given more weightage. The coding may accordingly

be done as 0 for most primitive character-state (simple leaf), 1 for intermediate character-state (pinnately lobed leaf) and 2 more most advanced state (pinnately compound leaf) (**Figure 8.15-I**). In molecular data, **transversions** (Purine to pyrimidine or pyrimidine to purine changes) are given more weightage (**Figure 8.15-V**) over **transitions** (purine to purine or pyrimidine to pyrimidine), because the latter occur more frequently and are easy to reverse, whereas the former is a less likely biochemical change. Restriction site gains may similarly be weighted over site losses. A complex character, presumably controlled by many genes, may change less easily than a simple character controlled by fewer genes. The former

is often given more weighting over a simple character. It may be assumed that leaf anatomy may not change easily but hairiness may change readily. The number of steps between two character states is conveniently represented through **character step matrix** (Figure 8.15). One may, however, be tempted to count leaf anatomy character as equivalent to two changes in hairiness. This may often be the result of bias to obtain desired results. It is reasonable, however, to adopt the approach of numerical taxonomy to give **equal weighting** to all the characters in the preliminary analysis, identify those characters which show the least homoplasy and give them more weightage in the subsequent analysis, a process known as **successive weighting**. This avoids a bias towards a particular character, and as such enables rational treatment of available data.

Residual weighting involves excluding a character from the list when information for a large number of taxa is not available, or is irrelevant. But in certain cases, information may be available for a particular character for large number of OTUs but not for a few. Alternately, the information may be irrelevant for a few taxa (say, the number of spurs in a taxon, which lacks spurs). Such characters are used in analysis but for the taxa for which information is not available or is irrelevant, an NC code (Not Comparable) is entered in the matrix. Whenever the NC code is encountered, the program bypasses that particular character for comparing the concerned taxon. For data handling by computers, the NC code is assigned a particular (not 0 or 1) numeric value. Such residual weighting should, however, be avoided when appreciable number of taxa are not comparable for a particular character. The coded data may be entered in the form of a matrix with **t** number of rows (OTUs) and **n** number of columns (character-states) with the dimension of the matrix (and the number of attributes) being **t** x **n** (Table 8.1).

Certain characters in plants evolve together. Occurrence of stipules and trilacunar nodes is usually **correlated**.

Similarly sympetalous members tend to have epipetaly and tenuinucellate ovules. Such **correlated characters** receive lesser weighting. If two characters are correlated, each gets 1/2 weighting, if three 1/3 weighting and so on.

It is always advisable to identify and include the most ancestral taxon (outgroup) as last taxon (or first taxon, as certain programs choose first taxon for rooting) in the list of taxa. If it is possible to identify plesiomorphic and apomorphic character-states, 0 represents plesiomorphic character-state and 1 the apomorphic character-state of a particular character. Outgroup taxon in the matrix gets 0 code for all character states (Table 8.4). For multistep changes, or unlikely events appropriate codes as indicated in Figure 8.15 are transferred to the matrix. Outgroup taxon in the matrix is essential in final rooting of the most parsimonious cladogram (tree).

Measure of similarity

Once the data have been codified and entered in the form of a matrix, the next step is to calculate the degree of resemblance between every pair of OTUs. A number of formulae have been proposed by various authors to calculate **similarity** or **dissimilarity (taxonomic distance)** between the OTUs. If we are calculating the similarity (or dissimilarity) based on binary data coded as 1 and 0, the following combinations are possible.

		OTU k	
		1	0
OTU j	1	<i>a</i>	<i>b</i>
	0	<i>c</i>	<i>d</i>

Number of matches $m = a + d$

Number of positive matches a

Number of mismatches $u = b + c$

Sample size $n = a + b + c + d = m + u$

j and **k** are two OTUs under comparison

Some of the common formulae are discussed below:

Simple matching coefficient

This measure of similarity is convenient and highly suitable for data wherein 0 and 1 represent two states of a character, and 0 does not merely represent the absence of a character-state. The coefficient was introduced by Sokal and Michener (1958). The coefficient is represented as:

$$S_{SM} = \frac{\text{Matches}}{\text{Matches} + \text{Mismatches}}$$

or

$$\frac{m}{m + u}$$

It is more convenient to record similarity in percentage (Table 8.2). In that case, the formula would read:

$$S_{SM} = \frac{m}{m + u} \times 100$$

When comparing a pair of OTUs, a match is scored when both OTUs show 1 or 0 for a particular character. On the other hand, if one OTU shows 0 and another 1 for a particular character, a mismatch is scored.

Jaccard Coefficient of association

The coefficient was first developed by Jaccard (1908) and gives weightage to scores of 1 only. This formula is thus suitable for data where absence-presence is coded and 1 represents the presence of a particular character-state, and 0 its absence. The formula is presented as:

$$S_j = \frac{a}{a + u}$$

where a stands for number of characters that are present (scored 1) in both OTUs. This can similarly be represented as a percentage similarity.

Yule coefficient

This coefficient has been less commonly used in numerical taxonomy. It is calculated as:

$$S_y = \frac{ac - bc}{ad - bc}$$

Taxonomic distance

Taxonomic distance between the OTUs can be easily calculated as a value 1 minus similarity or 100 minus percentage similarity. It can also be directly calculated as **Euclidean distance** using formula proposed by Sokal (1961):

$$\Delta_{jk} = \left[\sum_{i=1}^n (X_{ij} - X_{ik})^2 \right]^{1/2}$$

The average distance would be represented as:

$$d_{jk} = \sqrt{\frac{\Delta_{jk}^2}{n}}$$

Other commonly used distance measures include **Mean character difference** (M.C.D.) proposed by Cain and Harrison (1958), **Manhattan metric distance coefficient** (Lance and Williams, 1967) and **Coefficient of divergence** (Clark, 1952).

Once the similarity or distance between every pair of taxa has been calculated, the data are presented in a second matrix with $t \times t$ dimensions where both rows and columns represent taxa (Table 8.2; Table 8.3). It must be noted that diagonal t value in the matrix represents self-comparison of taxa and thus 100% similarity. These values are redundant as such. The values in the triangle above this diagonal line would be similar to the triangle below. The effective number of similarity values as such would be $t \times (t-1)/2$. Thus if 15 OTUs are compared the number of values calculated would be $15 \times (15-1)/2 = 105$.

A data matrix with coded character-states for each taxon can be used for calculating

Table 8.2 Similarity matrix of the representative hypothetical taxa presented as percentage simple matching coefficient.

OTUs	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	100														
2	47.0	100													
3	54.0	47.0	100												
4	49.0	54.0	52.0	100											
5	50.0	51.0	44.0	48.5	100										
6	46.0	59.0	46.0	47.0	48.0	100									
7	47.0	48.0	48.0	46.0	65.0	47.0	100								
8	56.0	51.0	56.0	51.5	46.0	58.0	25.0	100							
9	50.0	45.0	49.0	50.0	60.0	40.0	79.0	30.0	100						
10	50.0	45.0	54.0	50.5	58.0	41.0	77.0	36.0	92.0	100					
11	53.0	54.0	49.0	45.5	65.0	51.0	92.0	31.0	75.0	73.0	100				
12	48.0	47.0	49.0	50.0	58.0	42.0	81.0	30.0	96.0	94.0	75.0	100			
13	47.0	44.0	49.0	49.5	59.0	44.0	68.0	41.0	81.0	83.0	62.0	81.0	100		
14	55.0	46.0	55.0	51.5	57.0	44.0	72.0	39.0	81.0	81.0	72.0	81.0	74.0	100	
15	56.0	45.0	57.0	53.0	54.0	44.0	67.0	40.0	78.0	72.0	67.0	74.0	67.0	87.0	100

Table 8.3 Dissimilarity matrix of the representative hypothetical taxa based on the similarity matrix in Table 8.2.

OTUs	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.0														
2	53.0	0.0													
3	46.0	53.0	0.0												
4	51.0	46.0	48.0	0.0											
5	50.0	49.0	56.0	51.5	0.0										
6	54.0	41.0	54.0	53.0	52.0	0.0									
7	53.0	52.0	52.0	54.0	35.0	53.0	0.0								
8	44.0	49.0	44.0	48.5	54.0	42.0	75.0	0.0							
9	50.0	55.0	51.0	50.0	40.0	60.0	21.0	70.0	0.0						
10	50.0	55.0	46.0	49.5	42.0	59.0	23.0	64.0	8.0	0.0					
11	47.0	46.0	51.0	54.5	35.0	49.0	8.0	69.0	25.0	27.0	0.0				
12	52.0	53.0	51.0	50.0	42.0	58.0	19.0	70.0	4.0	6.0	25.0	0.0			
13	53.0	56.0	51.0	50.5	41.0	56.0	32.0	59.0	19.0	17.0	38.0	19.0	0.0		
14	45.0	54.0	45.0	48.5	43.0	56.0	28.0	61.0	19.0	19.0	28.0	19.0	26.0	0.0	
15	44.0	55.0	43.0	47.0	46.0	56.0	33.0	60.0	22.0	28.0	33.0	26.0	33.0	13.0	0.0

the distance (and, consequently, the similarity) between every pair of taxa, including the hypothetical ancestor. The distance is calculated as the total number of character-state differences between two concerned taxa, the data presented as **t x t** matrix (Table 8.5).

This method is closer to taxometric methods, because both plesiomorphic and apomorphic character-states are given equal weightage, but the inclusion of hypothetical ancestor is always crucial for the study.

Another method of calculating distance involves calculation of the number of apomorphic character-states common between the pairs of concerned taxa, ignoring the possession of plesiomorphic character-states in common (Table 8.6). Since only synapomorphy is likely to define monophyletic groups, this method is closer to the original cladistic concept.

Construction of Trees

Different methods are available for the final analysis of cladistic information. Three of these commonly used in phylogenetic analysis include Parsimony-based methods, Distance methods and Maximum likelihood method.

Parsimony-based methods

The methods are largely based on the biological principle that mutations are rare events. The methods attempt to minimise the number of mutations that a phylogenetic tree must invoke for all taxa under consideration. A tree that invokes minimum number of mutations (changes) is considered to be the tree of **maximum parsimony**. The evolutionary polarity of taxa is decided for construction of such trees. The **Wagner groundplan divergence method**, an example of this, was first developed by H. W. Wagner in 1948 as a technique for determining the phylogenetic relationships among organisms that he hoped would

replace intuition with analysis. The method was based on determining the apomorphic character-states present within a taxon and then linking the subtaxa based on relative degree of apomorphy. Interestingly, whereas the method found little favour with zoologists, it has been used in many botanical studies. Kluge and Farris (1969) and Farris (1970) developed a comprehensive methodology for the development of Wagner trees, based on the principle of parsimony. The method is the basis of many phylogeny computer algorithms currently in use. A given dataset may, however, yield many possible equally parsimonious trees due to homoplasy, as more than one character-state change may occur during the evolutionary process of a particular group of organisms.

The following steps are involved in the analysis:

1. Determine which of the various characters (or character-states) in a series of character transformations are apomorphic.
2. Assign the score of 0 to the plesiomorphic character and 1 to the apomorphic character in each transformation series. If the transformation series contains more than two homologues, then these 'intermediate apomorphies' may be scaled between 0 and 1. Thus, a transformation series of three characters may be scored as 0, 0.5 and 1 (or 0, 1 and 2 depending on the weighting assigned).
3. Construct a table of taxa (EUs) and coded characters (or character-states: see Table 8.3).
4. Determine the **divergence index** for each taxon by totalling up the values. Since apomorphic character-states are coded 1, the divergence index in effect represents the number of apomorphies (character-states) in a taxon, except in cases of weighted coding. For the data matrix in Table 8.4, the divergence index for 15 taxa would be calculated as:

Table 8.4 Data matrix of **t** taxa and **n** characters scored as 0 (plesiomorphic) and 1 (apomorphic) character-states. Multistate character is assigned 0 for ancestral state, 1 for intermediate and 2 for most advanced state. The matrix is similar to Table 9.1 but only 9 characters pictured are used for calculations. Also the last taxon included is the hypothetical ancestor in which all character-states are scored as 0 (plesiomorphic), as it is presumed that the ancestor would possess all characters in a plesiomorphic state.

Characters (n)→	Habit 0-woody 1-herbaceous	Fruit 0-follicle 1-achene	Ovary 0-superior 1-inferior	Leaves 0-simple 1-lobed 2-compound	Habitat 0-terrestrial 1-aquatic	Pollen 1-triporate 0-monosulcate	Ovule 1-unitegmic 0-bitegmic	Carpels 0-free 1-united	Plastids 1-PI-type 0-PII-type
Taxa (t) ↓									
1.	1	0	1	1	0	1	1	1	1
2.	1	0	1	1	1	1	0	0	1
3.	0	1	0	1	0	0	1	1	1
4.	1	1	1	0	1	0	0	0	0
5.	1	0	1	2	0	1	1	0	1
6.	1	1	0	1	1	1	0	1	0
7.	0	1	1	0	1	1	0	0	1
8.	0	0	0	1	0	0	1	1	1
9.	1	1	1	0	0	1	1	0	0
10.	0	0	0	1	1	0	1	1	1
11.	1	0	1	2	0	1	0	0	1
12.	1	1	0	0	1	1	0	0	1
13.	0	0	1	0	1	0	1	0	1
14.	1	0	1	0	1	0	1	0	1
15.	0	0	0	0	0	0	0	0	0

Taxon	Divergence index
1	7
2	6
3	5
4	4
5	7
6	6
7	5
8	4
9	5
10	5
11	6
12	5
13	4
14	5
15	0

Note that the hypothetical ancestral taxon 15 has an index of 0.

- Plot the taxa on a graph, placing each taxon on a concentric semicircle that

equals its divergence index. The lines connecting the taxa are determined by shared synapomorphies (see Table 8.6). The cladogram (**Wagner tree**) is presented in Figure 8.16.

Not all cladistic methods apply the principle of parsimony. The methods of **compatibility analysis** or **clique analysis** utilize the concept of character compatibility. Such methods can detect and thus omit homoplasy. They can be carried out manually or using a computer program, and can generate both rooted as well as unrooted trees. Groups of mutually compatible characters are termed **cliques**. Let us consider two characters, A and B, with two character-states each. Four character-state combinations are possible:

Assuming the evolution has proceeded from A1 to A2 and from B1 to B2. If all the

four combinations are met in nature then obviously there must have been at least one reversal (A2 to A1) or parallelism (A1 to A2 occurring twice), and as such A and B are incompatible. On the other hand, if only two or three of the combinations occur, then A and B are compatible. Cliques are formed by comparing all pairs of characters and finding mutually compatible sets. The largest clique is selected from the data to produce a cladogram. Finally, a rooted tree or network is obtained according to whether or not a hypothetical ancestor was included in the analysis.

Multiple Trees

The **unrooted tree** constructed in Figure 8.13-B represented only a small portion of the evolutionary sequence. Extension of this tree would make it more complicated and present a lot of possibilities. Let us add a

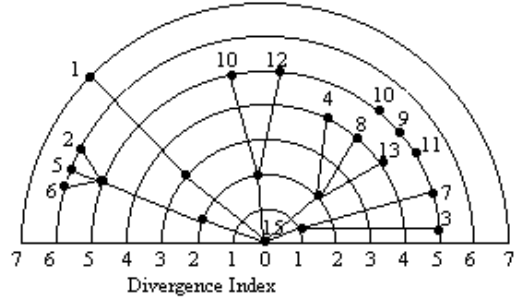


Fig. 8.16 General representation of a Wagner tree.

small portion of the herbaceous lineage with yellow petals, again assuming that there are a total of 15 herbaceous species of which six are with red petals and 9 with yellow petals. Of these nine 4 are with united carpels and 5 with free carpels. The additional Venn dia-

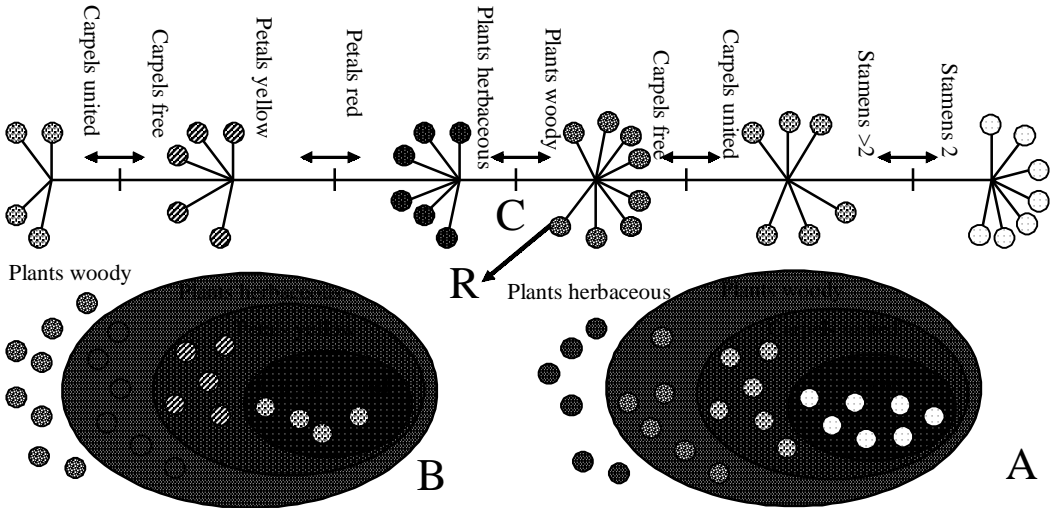


Figure 8.17 **A:** The Venn diagram for woody species, the same as Figure 10.12A. **B:** The Venn diagram for a small portion of the herbaceous lineage of assumed 15 species of which 6 are with red petals and 9 with yellow petals, latter with 4 species having united carpels and 5 free carpels. **C:** Extension of the unrooted tree of Figure 10.12B to include the species depicted in the Venn diagram B here. There are 5 actual character state changes but with 4 switches as united carpels have arisen twice.

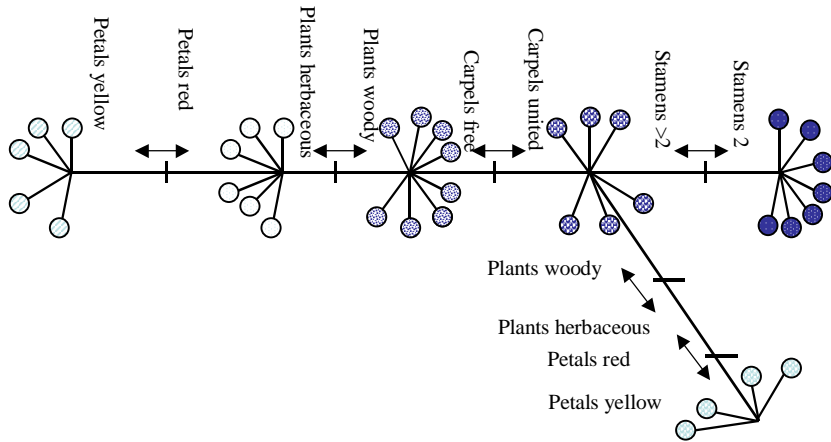


Figure 8.18 Possible variation of the unrooted tree presented in figure 10.14C, if we did not have any idea about the evolutionary history of the group. Note that tree length has increased to six, and habit has changed twice from woody to herbaceous and from red to yellow petals. Such homoplasious situations are uncommon.

gram for the herbaceous species and the extended unrooted tree is presented in the Figure 8.17-C. It must be noted that here we know the evolutionary history of the group—which normally is never known—and the aim of phylogenetic analysis is to reconstruct and depict this evolutionary history through trees. The unrooted tree here has five character-state changes (actual changes, tree length) involved. The change from free to united carpels has occurred twice, and as such there are only four genetic switches involved. If we did not have the knowledge about the evolutionary history of the group, we would try a number of variations. One possible variation of the unrooted tree would be to link 4 herbaceous species with united carpels to the woody species with united carpels, thus presenting a single change of free to united carpels. But this brings in further changes. Now, change from woody habit has occurred twice, change from red to yellow petals has occurred twice, and more significantly the number of actual changes (tree length) has increased to six (Figure 8.18), with same four genetic switches involved. With more descendants

being included in the tree, the number of options would increase. Also we have to convert each unrooted tree into a **rooted tree** so that the most primitive basal end of the tree is known, and different lineages presented as the more advanced branches. This brings in many more options, as indicated earlier. In our example, where we know the history of the tree, the tree can be rooted at R, as indicated by an arrow (Figure 8.17-C), but in a large majority of cases, it is a complicated process, and a lot of hypotheses, strategies and algorithms come into play.

A number of sophisticated computer algorithms are available which compare trees and calculate their lengths. The widely used ones include NONA, PAUP, and PHYLIP. These programs determine the number of possible trees, and then sort out the shortest of all these. If we are dealing with three species, three rooted trees are possible [A (B, C)], [B (A, C)] and [C (A, B)] (Figure 8.19), if 4 taxa are mapped the 15 trees are possible, 5 then 105, for 10 taxa 34,459,425 trees and so on.

The number of possible rooted trees for n number of taxa can be calculated as:

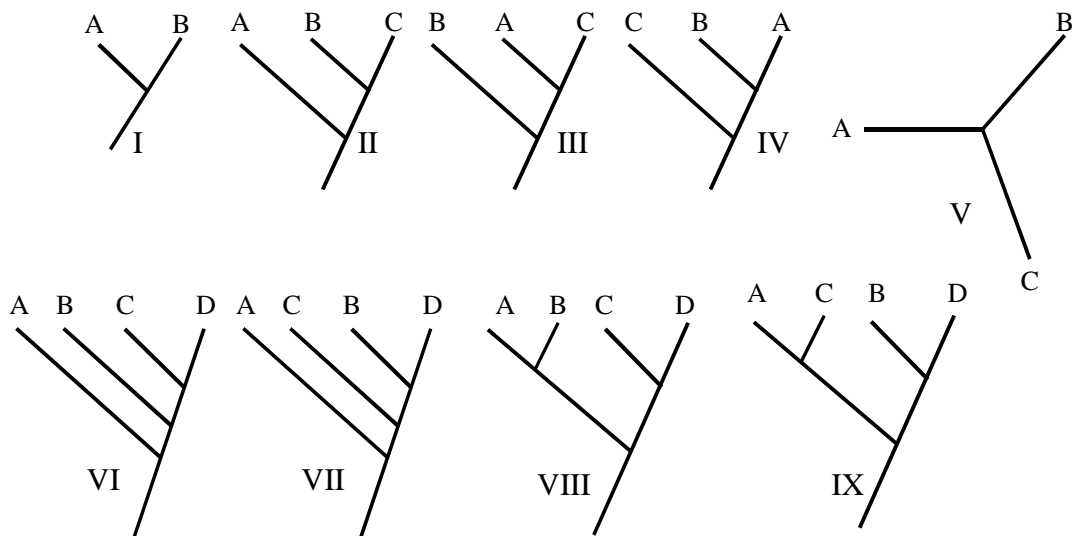


Figure 8.19 Possible number of rooted and unrooted trees. I: Single rooted tree for two taxa. II-IV: Three possible rooted trees for three taxa. V: One possible unrooted tree for three taxa. VI-IX: Some of the possible 15 rooted trees for four taxa.

$$Nr = (2n-3)! / [(2^{n-2}) \times (n-2)!]$$

It can also be calculated as:

$$Nr = P(2i-1)$$

where P represents the product of all factors $(2i-1)$ from $i=1$ to $i=n-1$.

A simpler way to calculate the possible number of rooted trees is as follows:

$$Nr = (2(n+1)-5) \times \text{number of trees for } (n-1) \text{ taxa}$$

As noted above, the number of possible rooted trees is much more than number of possible unrooted trees. Latter can be calculated as:

$$Nu = (2n-5)! / [(2^{n-3}) \times (n-3)!]$$

or more simply as:

Nu = number of rooted trees for $(n-1)$ taxa

Thus, for 3 taxa, 3 rooted trees and 1 unrooted trees are possible (Figure 8.19), for 4 taxa 15 rooted trees and 3 unrooted trees and for 5 taxa, 105 rooted trees are possible but only 15 unrooted trees. For our 8 species in Figure 8.12, if evolutionary history

was not known we should expect 135135 rooted trees and 10395 unrooted trees. The figures also highlight the enormous challenges in reconstructing the evolutionary history of any group.

A large number of trees generated are sorted and, ones presenting the shortest evolutionary path, in agreement with the **principle of parsimony**, are shortlisted.

Distance methods

Distance methods were originally developed for handling phenetic information and construction of phenograms, some of these have now been incorporated in cladistic methodology. Cluster analysis is the most commonly used method of constructing trees.

Cluster analysis

Data presented in OTUs \times OTUs ($t \times t$) matrix are too exhaustive to provide any meaningful picture and need to be further condensed to enable a comparison of units. Cluster analysis is one such method in which OTUs are arranged in the in the

order of decreasing similarity. The earlier methods of cluster analysis were cumbersome and involved shifting of cells with similar values in the matrix so that OTUs with closely similar similarity values were brought together as clusters. Today, with the advancement of computer technology, programs are available which can perform an efficient cluster analysis and help in the construction of cluster diagrams or phenograms. The various clustering procedures are classified under two categories.

Agglomerative methods

Agglomerative methods start with t clusters equal to the number of OTUs. These are successively merged until a single cluster has finally been formed. The most commonly used clustering method in biology is the **Sequential Agglomerative Hierarchic Non-overlapping clustering method** (SAHN). The method is useful for achieving hierarchical classifications. The procedure starts with the assumption that only those OTUs would be merged which show 100% similarity. As no two OTUs would show 100% similarity, we start with t number of clusters. Let us now lower the criterion for merger as 99% similarity; still no OTUs would be merged as in our example the highest similarity recorded is 96.0%. The best logical solution would be to pick up the highest similarity value (here 96.0) and merge the two concerned OTUs (here 9 and 12). By inference, if our criterion for merger is 96.0 we will have $t-1$ clusters. Subsequently the next lower similarity value is picked up and the number of clusters reduced to $t-2$. The procedure is continued until we are left with a single cluster at the lowest significant similarity value. Since at various steps of clustering a candidate OTU for merger would cluster with a group of OTUs, it is important to decide the value that would link the clusters horizontally in a cluster diagram. A number of strategies are used for the purpose.

In the commonly used **single linkage clustering method** (**nearest neighbour**

technique or **minimum method**), the candidate OTU for admission to a cluster has similarity to that cluster equal to the similarity to the closest member within the cluster. The connections between OTUs and clusters and between two clusters are established by single links between pairs of OTUs. This procedure frequently leads to long straggly clusters in comparison with other SAHN cluster methods. The phenogram for our data using this strategy is shown in Figure 8.20.

The highest similarity value in our matrix (see Table 8.2) is 96.0 between OTUs 9 and 12, and as such they are linked at that level. The next similarity value of 94.0 is between OTUs 10 and 12, but since 12 has already been clustered with 9, 10 will join this cluster linked at 94.0. The process is repeated till all OTUs have been agglomerated into single cluster at similarity value of 53.0.

In the **complete linkage clustering method** (**farthest neighbour** or **maximum method**) the candidate OTU for admission to a cluster has similarity to that cluster equal to its similarity to the farthest member within the cluster. This method will generally lead to tight discrete clusters that join others only with difficulty and at relatively low overall similarity values.

In the **average linkage clustering method**, an average of similarity is calculated between a candidate OTU and a cluster or between two clusters. Several variations of this average method are used. The unweighted pair-group method using arithmetic averages (**UPGMA**) computes the average similarity or dissimilarity of a candidate OTU to a cluster, weighting each OTU in the cluster equally, regardless of its structural subdivision. The method originally developed for the procedures of numerical taxonomy has been applied in phylogenetic analysis with relevant modifications, and used for the construction of trees. **UPGMA method** procedure begins with as many clusters as the number of taxa. The two taxa with **minimum distance** merged to reduce the number of clusters by one. In the next

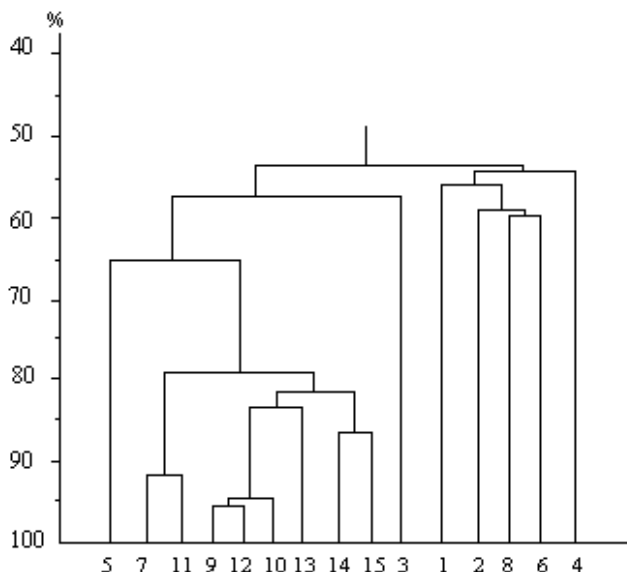


Figure 8.20 Cluster diagram of 15 OTUs based on similarity matrix in Table 8.2 using single linkage strategy.

step average distance between new cluster and remaining taxa are determined by taking the average distance between these two members and all other remaining taxa, weighting each taxon in the cluster equally regardless of its structural subdivision, and merging the taxon with smallest distance to the first cluster. The process is repeated with this new cluster of three taxa, and the procedure continues till all the taxa are merged, the most distant taxon joining last of all. From measure of similarity or dissimilarity of taxa (OEUs) as presented in Table 8.5 and 8.6, a network presenting minimum dissimilarity is constructed. Analysis of data from first six taxa of table 8.4 is presented in Figure 8.21. The procedure begins by uniting nearest taxa A and E (with minimum distance of 2). Next matrix is now constructed in which distance between (AE) and rest of the taxa is recalculated. The lowest value in this matrix (step 1 matrix) is between (AE) and E, which are next united at distance level 3 into (AE)B. The distance between this cluster and rest of the taxa is

now recalculated as presented in step 2 matrix. The lowest distance in this matrix is 4 between D and F which are united into one cluster. With this merger the distance between taxa/clusters is recalculated and presented in step 3 matrix. The lowest distance 5.5 is now between clusters (AE)B and DF, which are next united. Finally the distance between this enlarged cluster and C is recalculated as presented in step 4 matrix. Finally the two clusters (((AE)B)(DF)) and C are united at distance of 6.5 to form final cluster (((AE)B)(DF))C. The resulting phenogram and reconstructed phylogenetic tree constructed from the analysis are presented in Figure 8.21.

The distance matrix can similarly be generated from single nucleotide differences between homologous DNA sequences derived from different species. Six such hypothetical sequences from different species are presented in Figure 8.22. The distance matrix is based on number nucleotide differences between different sequences. A and F with lowest distance are merged, followed

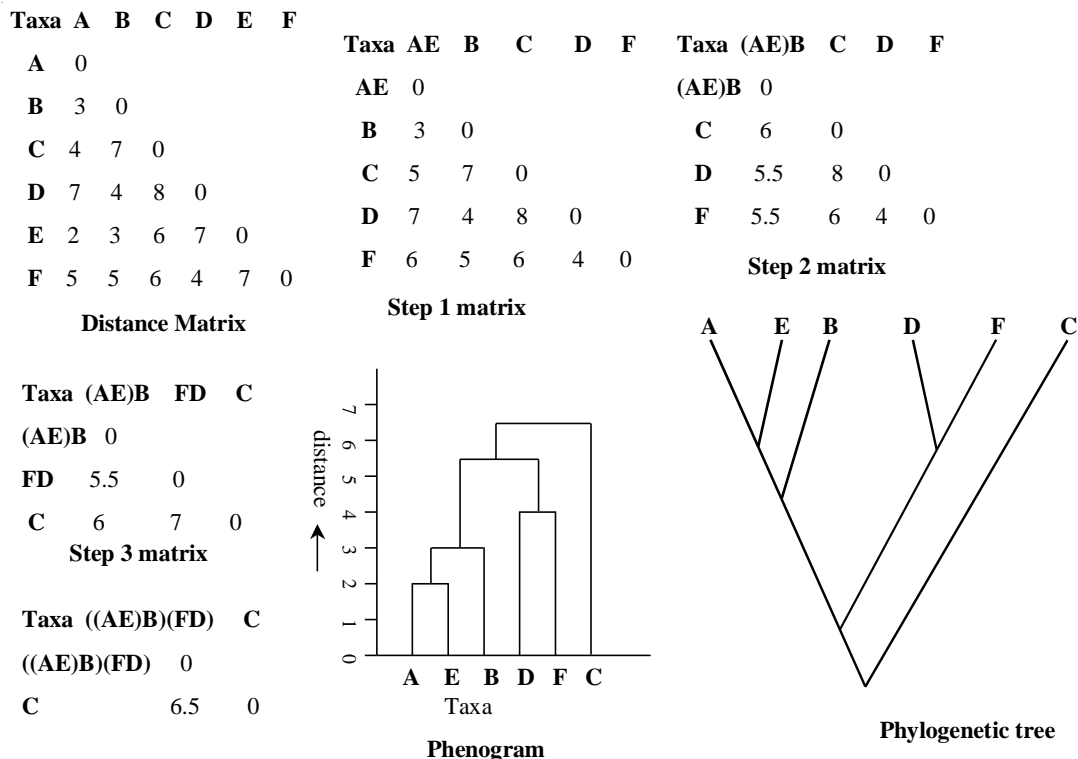


Figure 8.21 Construction of phenogram and phylogenetic tree (cladogram) based on distance matrix concerning first six taxa in Table 8.4 using UPGMA clustering method. The taxa with minimum distance are united and treated as single cluster in next matrix, and distance values recalculated as average of distance from either of united taxa. The procedure repeated till all taxa are united. The phylogenetic tree is constructed based on sequence of clustering of taxa.

by recalculation of new matrix in which A and F form one cluster and value of each taxon is calculated as average distance from A and F. Now lowest value is shared by B and D which form second cluster. The values are recalculated similarly, and successively C joins AF cluster, and then E joins (AF)C cluster. The two clusters are finally merged to enable construction of phylogenetic tree either as phenogram or as cladogram. Some types of genetic polymorphism data such as RAPD are best handled when sharing of 0 code by two taxa in the matrix is ignored when both taxa lack a given polymorphic band in gel electrophoresis. Jaccard coefficient is best suited for handling such data.

Figure 8.23 presents results of RAPD analysis of 8 taxa, where only polymorphic bands are shown, monomorphic bands being omitted. The distance matrix based on similarity matrix was processed using PHYLIP, as shown in Figure 8.24. Distance methods are suitable for handling both morphological and molecular data, or a combination of both. These methods use all data with usually equal importance, whereas the parsimony methods use only informative molecular data. In general for a site to be informative, when handling sequence data, irrespective of how many sequences are aligned, it has to have at least two different nucleotides, and each of these nucleotides has to be

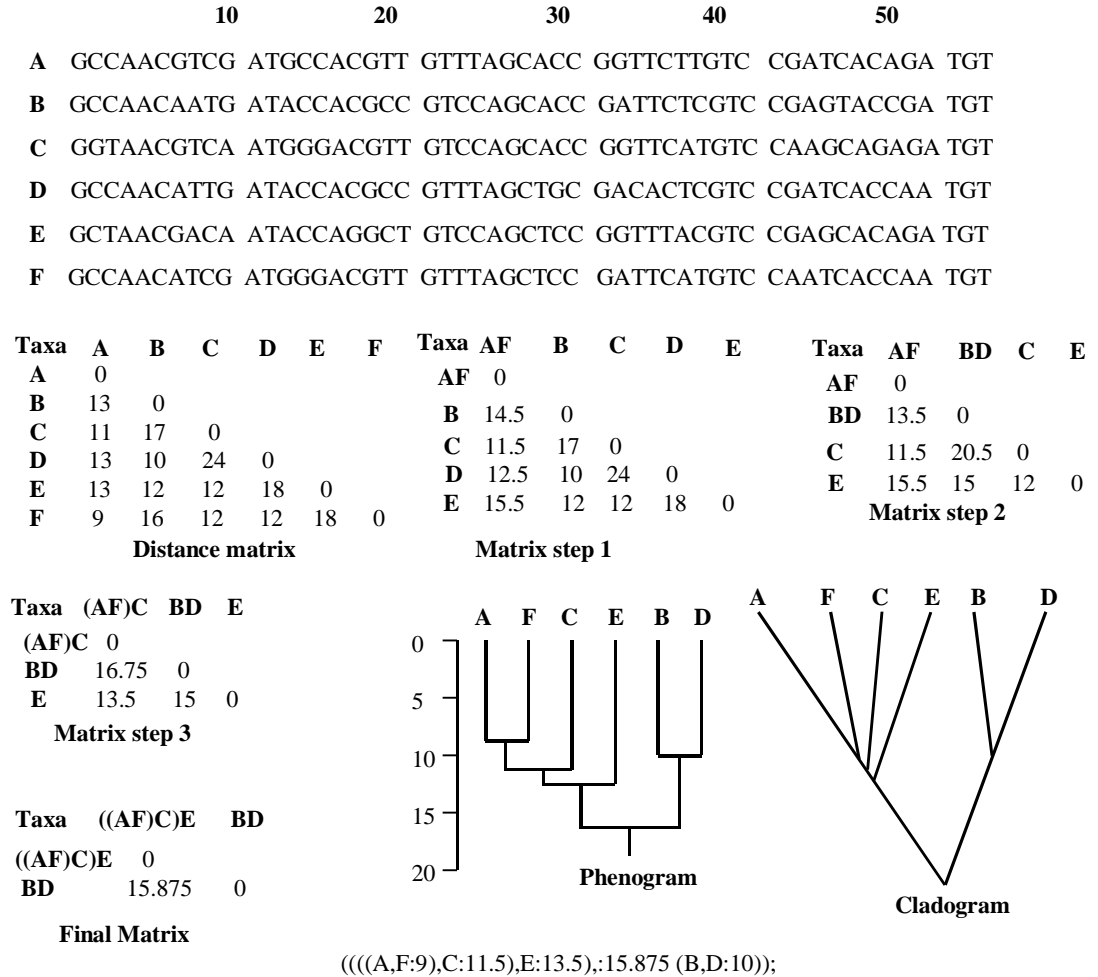


Figure 8.22 Construction of phylogenetic tree based on single nucleotide differences in hypothetical DNA sequences of six species. Distance matrix is constructed based on the number of nucleotide differences between each pair of DNA sequences and presented in distance matrix. Further analysis proceeds as detailed in Figure 8.21, and also in the text on these pages.

present at least twice. Thus in the sequence data presented in Figure 8.22, out of 28 sites showing nucleotide differences in six sequences, there are only 12 informative sites which can be used in parsimony analysis. Procedures based on UPGMA method, however, don't account for different rates of evo-

lution occurring in different lineages. Some distance methods such as **transformed distance method** and **neighbour-joining method**, although more complex are capable of incorporating different rates of evolution within the lineages.

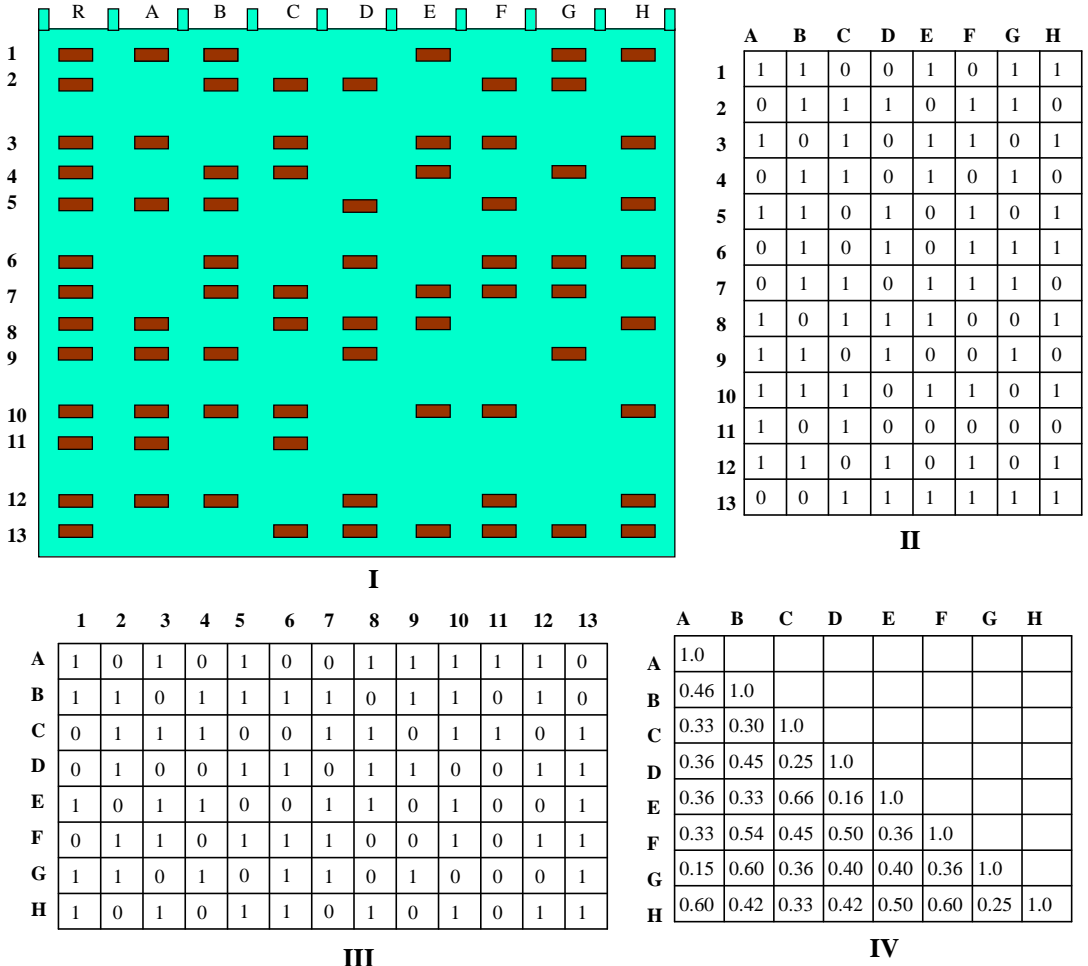


Figure 8.23 Phylogenetic analysis of data concerning polymorphic bands from gel electrophoresis from DNA of 8 taxa. **I**: Polymorphic bands of DNA of 8 taxa (A-H), R representing reference bands; **II**: Binary coded matrix of the polymorphic bands; **III**: The same matrix presented in conventional format; **IV**: Lower triangular matrix of similarity matrix using Jaccard coefficient, wherein sharing of 0 state (absence of bands) is ignored. Further handling of data using UPGMA program of PHYLIP is presented in Figure 8.24.

Divisive methods

Divisive methods as opposed to agglomerative methods, start with all *t* OTUs as a single set, subdividing this into one or more subsets; this is continued until further subdivision is not necessary. The commonly used divisive method is **associa-**

tion analysis (William, Lambert and Lance, 1966). The method has been mostly used in ecological data employing two state characters. It builds a dendrogram from the top downwards as opposed to cluster analysis, which builds a diagram from the bottom up. The first step in the analysis involves

8

Taxona	0.00	0.54	0.67	0.64	0.64	0.67	0.85	0.40
Taxonb	0.54	0.00	0.70	0.55	0.67	0.46	0.40	0.58
Taxonc	0.67	0.70	0.00	0.75	0.34	0.55	0.64	0.67
Taxond	0.64	0.55	0.75	0.00	0.84	0.50	0.60	0.58
Taxone	0.64	0.67	0.34	0.84	0.00	0.64	0.60	0.50
Taxonf	0.67	0.46	0.55	0.50	0.64	0.00	0.64	0.40
Taxong	0.85	0.40	0.64	0.60	0.60	0.64	0.00	0.75
Taxonh	0.40	0.58	0.67	0.58	0.50	0.40	0.75	0.00

I

((Taxona:0.20000,Taxonh:0.20000):0.11000,(Taxonc:0.17000,

Taxone:0.17000):0.14000):0.01500,((Taxonb:0.20000,Taxong:0.20000):0.08125,

II

(Taxond:0.25000,Taxonf:0.25000):0.03125):0.04375);

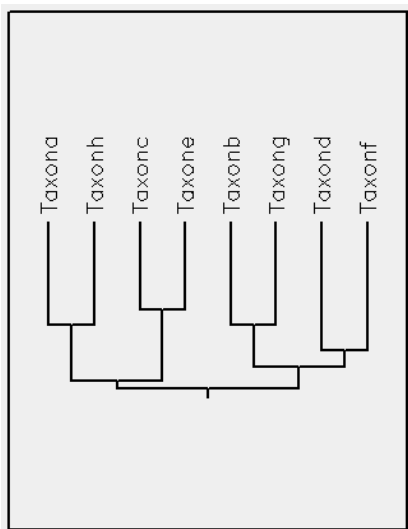
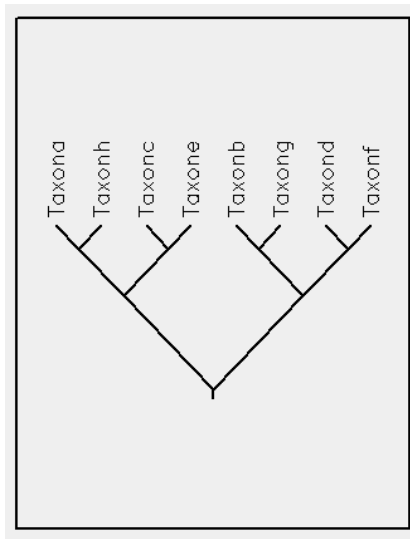
**III****IV**

Figure 8.24 Construction of phylogenetic tree based on polymorphic bands from gel electrophoresis from DNA of 8 taxa using UPGMA program of PHYLIP; **I**: Square distance generated from Figure 8.23-IV, each value calculated 1-similarity value. **II**: Outtree file generated by UPGMA option of NEIGHBOUR program; **III**: Upright square tree (Phenogram) plotted through DRAWGRAM program; **IV**: Cladogram, but with branch lengths omitted.

calculating **chi square** value between every pair of characters using the formula:

$$X_{hi}^2 = \frac{n(ad - bc)^2}{[(a + b)(a + c)(b + d)(c + d)]}$$

where i stand for the character being compared and h for any character other than i .

For each character the sum of chi-square is computed and the character showing maximum **chi square** value is chosen as the first differentiating character. The whole set of OTUs is divided into two clusters, one containing the OTUs which show the character-state a and another containing OTUs which show the character-state b . Within

each cluster, again, the character with the next value of the sum of chi square is selected and the cluster subdivided into two clusters as before. The process is repeated till further subdivision is not significant.

Hierarchical classifications

The phenogram constructed using any technique or strategy can be used for attempting hierarchical classification, by deciding about certain threshold levels for different ranks. One may tentatively decide 85 per cent similarity as the threshold for the species, 65 for genera and 45 for families and recognize these ranks on the basis of number of clusters established at that threshold. Whereas such an assumption can help in hierarchical classification, the point of conflict would always be the threshold level for a particular rank. Some may argue—and are justified in doing so—to suggest 80 per cent (or any other value) as the threshold for species. It is more common, therefore, to use terms **85 per cent phenon line, 65 per cent phenon line, and 45 per cent phenon lines**. These terms may conveniently be used till such time that sufficient data are available to assign them formal taxonomic ranks to the various phenon lines.

The results of cluster analysis are commonly presented as dendrograms known as **phenograms**. They can also be presented as **contour diagrams** (Figure 8.25), originally developed under the name **Wroclaw diagram** by Polish phytosociologists. The contour diagram may also incorporate the levels at which clustering has taken place.

Ordination

Ordination is a technique which determines the placement of OTUs in two-dimensional or three-dimensional space. The results of two-dimensional ordination are conveniently represented with the help of a scatter diagram and those of three-dimensional ordination with the help of a three-dimensional model. The procedure works on distance values calculated directly from the coded data or indirectly from the already cal-

culated similarity values as 100 minus similarity (if similarity values are in percentage) or 1 minus similarity (if similarity values range between 0 to 1). A dissimilarity matrix based on Table 8.2 is presented in Table 8.3.

The first step in the ordination starts with construction of the x-axis (horizontal axis). In the commonly used method of **polar ordination**, the two most distant OTUs are selected as the end points (A and B) on x-axis. In our example, these are OTU 8 and 7 with a distance (dissimilarity value) of 75. The position of all other OTUs on this axis can be plotted one by one. OTU 10 has a distance of 64 from A (OTU 8) and a distance of 23 from B (OTU 7). A compass with a radius of 64 units is swung from A and a compass with a radius of 23 units is swung from B, forming two arcs. A line joining the intersection of two arcs forms a perpendicular on the x-axis, and the point at which the line crosses the x-axis is the position of the OTU. The distance between the x-axis and the point of intersection of arcs is the **poorness of fit** of the concerned OTU. The location of OTU on the axis from the left (point A) can also be calculated directly instead of plotting:

$$x = \frac{L^2 + dAC^2 - dBC^2}{2L}$$

where x is the distance from the left end, L is the dissimilarity value between A and B (length of x-axis), dAC is dissimilarity between A and the OTU under consideration and dBC as the dissimilarity between B and the OTU under consideration. The poorness of fit (e) of this OTU can be calculated as:

$$e = \sqrt{dAC^2 - x^2}$$

After the position of all OTUs has been determined and the poorness of fit calculated, a second axis (vertical axis or y-axis) has to be calculated. For this, the OTU with the highest poorness of fit (most poorly fitted to x-axis) is selected and this forms the first reference OTU of y-axis. The second reference OTU is selected as that one with the

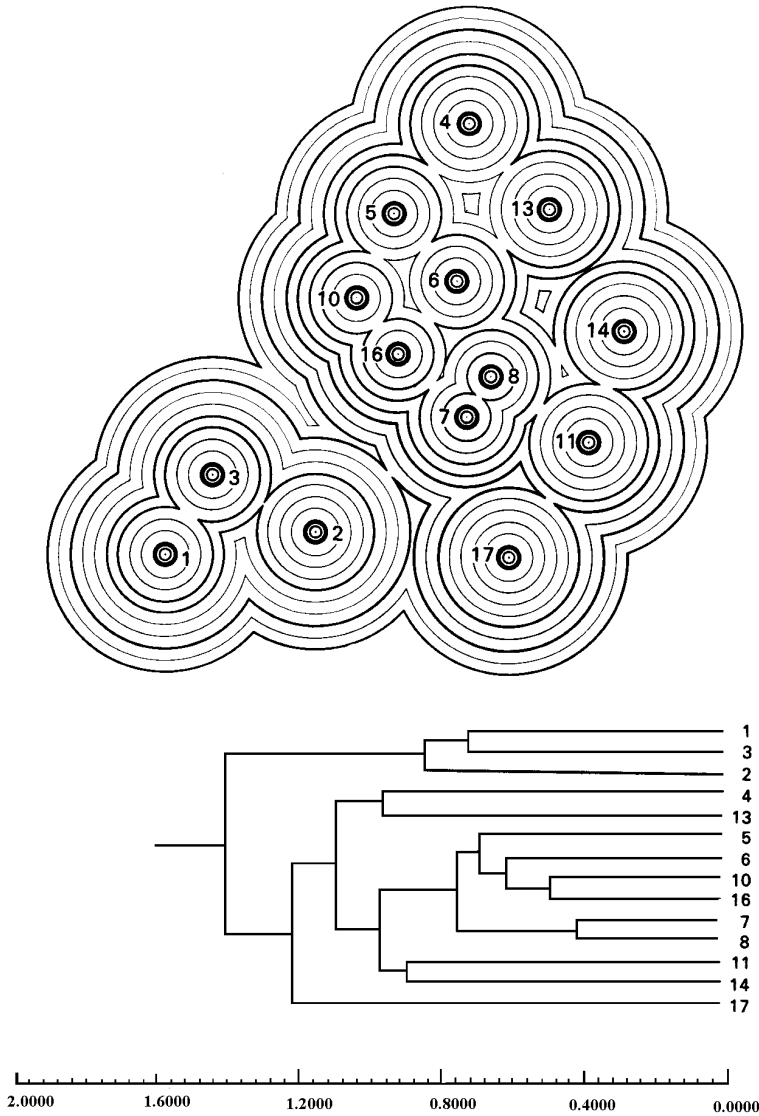


Figure 8.25 Contour diagram based on the phenogram shown alongside.

highest dissimilarity to the first reference OTU of y-axis, but within 10 per cent (of the length of x-axis) distance on x-axis. The position of all other OTUs on the y-axis and their poorness of fit is determined as earlier. By using the values of poorness of fit to y-axis, a z-axis can be similarly generated and the position of all OTUs on z-axis determined similarly. The values can be used for

constructing a **scatter diagram** or a three-dimensional model.

A commonly used ordination technique known as **principal component analysis** also calculates values for a two-dimensional scatter diagram. In this method, however, the values on the horizontal as well as the vertical axis are non-zero, ranging from -1 to 1 (calculated as **eigenvalues**) and as such the

scatter diagram is presented along four axes: positive horizontal, negative horizontal, positive vertical and negative vertical (Fig 8.26). The technique is based on the assumption that if a straight line represented a single character, all the OTUs could be placed along the line according to their value for that character. If two characters were used, a two-dimensional graph would suffice to locate all OTUs. With n characters, n -dimensional space is required to locate all OTUs as points in space.

Principal component analysis determines the line through the cloud of points that accounts for the greatest amount of variation. This is the first principal component axis. A second axis, produced perpendicular to the first, accounts for the next greatest amount of variation. The procedure ultimately produces axes one less than the number of OTUs. The first two axes are generally plotted to produce a scatter diagram. The procedure also calculates **eigenvectors**, which indicate the importance of a character to a particular axis. The larger the eigenvector in absolute value, the more important is that particular character.

A related method of ordination is **principal co-ordinate analysis** developed by Gower (1966). This technique enables computation of principal components of any Euclidean distance matrix without being in possession of original data matrix. The method is also applicable to non-Euclidean distance and association coefficients as long as the matrix has no large negative eigenvalues. Principal co-ordinate analysis also seems to be less disturbed by NC entries than principal components.

Maximum Likelihood method

The method is similar to distant method in that all data is taken into consideration. In this method, similarly character-state transformations are compared, and the probability of changes determined. These probabilities are used to calculate the likelihood that a given tree would lead to the particular data set observed, and the tree with maximum

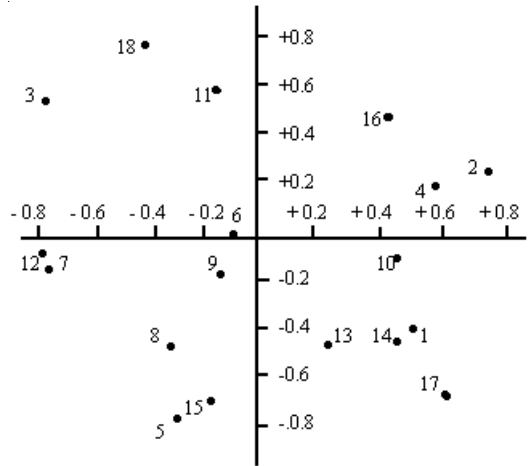


Figure 8.26 Plot of the results of the principal component analysis of 18 hypothetical taxa.

likelihood is selected. The method is especially suited to molecular data, where the probability of genetic changes can be modeled more easily. With this approach, the probabilities are considered for every individual nucleotide substitution in a set of sequence alignments. It is commonly understood that transitions occur three times more frequently as compared to transversions. Thus if C, T, and A occur in one column (representing one site), the sequences with C and T (pyrimidines) are more likely to be closely related than sequence with A (Purine). Using objective criteria probability for each site and every possible tree that describes the relationship of sequences. The tree with highest aggregate probability is selected as representation of a true phylogenetic tree.

Using any one of the methods, a large dataset commonly used, and which includes many homoplasies, large number of shortest trees may be generated by these automated algorithms. These short listed trees have to be further compared.

The Consensus Tree

The use of automated methods based on parsimony, even after applying relevant strate-

gies yield several trees, all presenting shortest pathways, based on parsimony but with different linkages among the taxa (OEU's), and often presenting different evolutionary history. Molecular studies of Clusiaceae by Gustafsson et al., (2002) for example, including 1409 nucleotides of chloroplast gene *rbcL* positions using PAUP*4.0b8a parsimony analysis method, yielded 8473 most parsimonious trees for the 26 species compared. Interestingly, the number of trees generated was so large that search for trees 3 steps longer than most parsimonious trees was aborted. More significantly different data sets (molecular, morphology) may yield different trees. While selecting the consensus tree, the commonest approach is to identify the groups, which are found in all the short listed trees, and build a **consensus tree**. This could be achieved in different ways.

Strict consensus tree

A more conservative approach in building a consensus tree involves including only monophyletic groups that are common to all the trees. The tree developed this way is known as **strict consensus tree**. Consider the two most parsimonious trees (although there could often be numerous trees of same shortest length available for comparison) as shown in Figure 8.27-I and 8.27-II.

Imagine that all groups A to J are monophyletic. Tree I shows that A and B are very closely related, and so are H and I. C, D, E, and F are shown arising successively and are related in that sequence. Tree II shows a similar relationship between H and I, and between A and B (but group J is shown related to these two). The tree also shows that C and D are closely related. As relationships between E, F, and G are ambiguous, they are shown arising from the same point in evolutionary history. The consensus tree III would thus omit taxon J (which is absent from tree I), show A and B, as also H and I as in the two trees I and II. The other taxa C, D, E, F, and G are shown arising from the same point.

Majority-rule consensus tree

Majority-rule consensus tree shows all the groups which appear in a majority of trees, say, more than 50 per cent of the trees. It is useful to indicate for each group on the consensus tree the percentage number of the most parsimonious trees in which the group appeared. Such a consensus tree, however, provides a partial summary of the phylogenetic analyses, and may be inconsistent with the trees from which it is derived.

Semi-strict consensus tree

A semi-strict consensus tree is useful when comparing trees from different data sets, or with different terminal taxa. The consensus tree developed indicates all the relationships supported by both type of trees or any one of these, but not contradicted by any. Thus, in Figure 8.27, tree II does not give us any information about the time of origin of E, F and G, the tree I indicates that they originated successively. Similarly, tree I does not indicate any close relationship between C and D, whereas the tree II does. The semi-strict consensus tree IV as such presents such information, not contradicted by either tree.

Evaluating consensus tree

Developing a consensus tree involves the use of intuition, making guesses and developing hypothesis. A number of evaluation strategies are used to test the soundness of the tree and measuring support for either the tree as a whole or for its individual branches. These values are generally published along with the tree, to allow the fair assessment of the final results for comparison of trees based on different datasets.

Consistency Index

The principle of Parsimony is based on a basic rule of science known as **Ockham's razor**, which says '**do not generate a hypothesis any more complex than is demanded by the data**'. Some information in the data may be representing homoplasy (reversals,

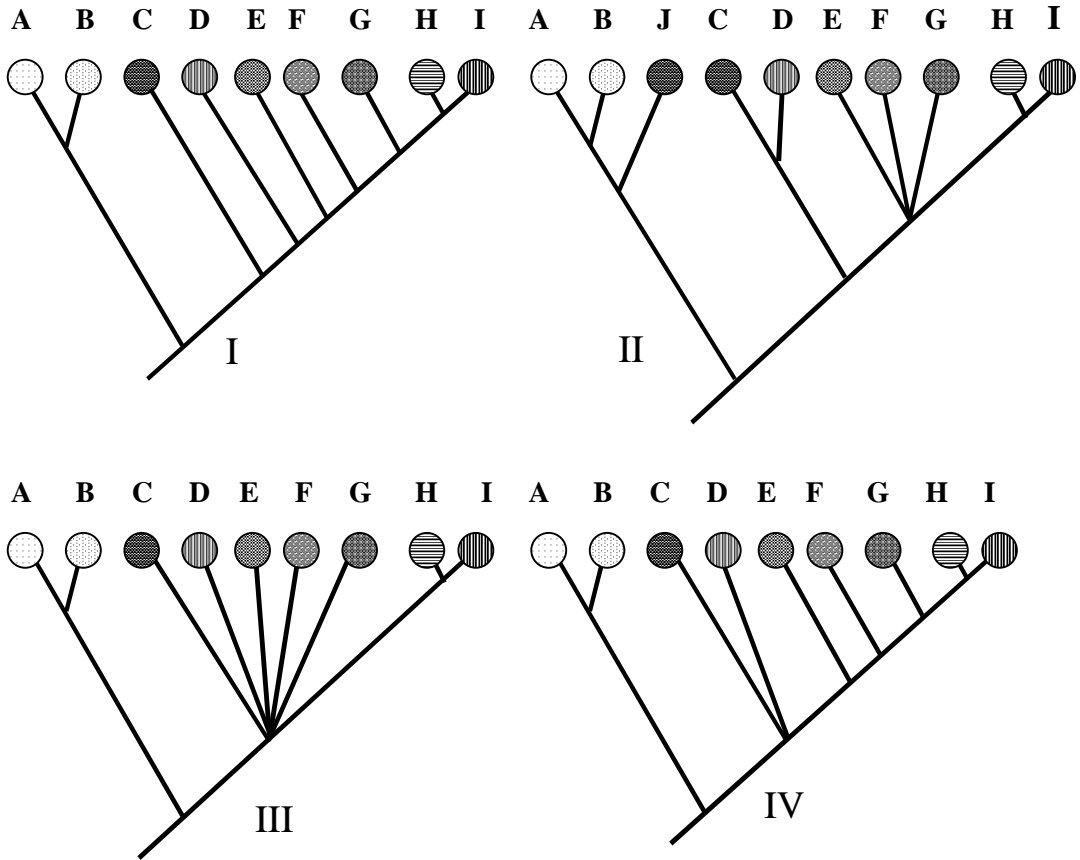


Figure 8.27 Two most parsimonious trees for a particular group of organisms, with monophyletic taxa A to J. **I:** showing C, D, E and F arising successively. **II:** E, F, G are shown arising at the same time from a common point, C and D being closely related. **III:** Strict consensus tree of trees I and II. **IV:** Semi-strict consensus tree of trees I and II.

parallelisms). Dollo parsimony (as indicated above) minimizes the use of homoplasious characters. The commonest measure of homoplasy is the **Consistency Index (CI)**, which is calculated by dividing the number of genetic switches by actual genetic changes on the tree.

$$\text{Consistency Index } CI = \text{Min} / L$$

Min stands for the minimum possible tree length or genetic switches, and *L* for the actual tree length or actual number of genetic

changes. In the tree shown in Figure 8.13B, there are three character-state changes, each involving one switch, and, as such, the consistency index would measure $3/3 = 1$. The tree shown in Figure 8.17C has five actual character-state changes (tree length is 5), but it involves only four genetic switches. As carpel fusion has occurred twice, the consistency index would accordingly be $4/5 = 0.8$. In the tree shown in Figure 8.18, the number of genetic switches remains the same as four but the tree length has increased to six due to two

parallel (or convergent) evolutions; the CI would be calculated as $4/6 = 0.66$.

Consistency Index may also be calculated for individual characters. In Figure 8.13B as such CI for all characters is one, while in 8.17-C, it is 0.5 for carpel fusion (minimum number of changes possible—one for binary character divided by actual number of changes—here 2 since the character has changed twice) and 1 for rest. In Figure 8.18, CI is 0.5 for habit and petal colour, and 1 for stamen number and carpel fusion. The characters that lower the CI of a tree (or which have lower CI) are considered to be homoplasious. The inclusion of a larger number of homoplasious characters in the analysis lowers CI for the tree and contradicts phylogeny. There may also be a character, which changes only in one (or a very few) species, and may be of no relevance in others. Suppose one species develops spiny fruits. The length or number of spines would not be of any relevance in rest of the species without spines. Such a situation (a single species having a particular character) is known as **autapomorphy**. Since such a character has changed only once, it gives CI of 1, and as such the inclusion of many such characters would increase the consistency index of the tree, and provide false support. Such uninformative characters are as such omitted before calculating CI.

The Consistency index values are often dependent on the number of taxa analyzed. Any increase in number of taxa lowers CI values, and this is true for data from different sources, morphological or molecular.

Retention Index

Although theoretically the value of CI could range between 0 and 1, it rarely goes below 0.5. For a character that, has changed five times on a tree (this is a remote possibility), CI will be 0.2. More so, the value of CI for a tree, very rarely may go below 0.5, and the values thus range between 0.5 and 1. The **Retention Index (RI)** corrects this narrow range of CI by comparing maximum (and not minimum as in CI) possible number of

changes in the character with actual number of changes in the character. RI is computed by first calculating the maximum possible tree length, if the apomorphic character-state originated independently in every taxon that it appears in, or say, the taxa are unrelated for the said character-state. The value of RI is calculated as:

$$\text{Retention Index } RI = (Max - L) / (Max - Min)$$

Max stands for the maximum tree length possible, *L* the actual tree length and *Min* the minimum tree length possible. The tree in Figure 8.17-C thus has a maximum possible tree length of 9 (minimum length of 4 and actual length of 5 as we already know) and the RI would be $(9-5)/(9-4) = 0.8$. Higher the RI, sounder is the tree.

Bremer Support (Decay Index)

The principle of parsimony, followed in phylogenetic analyses, aims at selecting the shortest tree. Some parts of the tree may be more reliable than others. This is commonly evaluated by comparing the shortest tree with those one or more steps longer. **Decay index** or **Bremer Support** is the measure of how many extra steps are needed before the original clade (group) is not retained. Thus if an internode has decay index of 3, then the clade (monophyletic group) arising from it is maintained even in the cladogram 3 steps longer than the shortest tree (see [Figure 8.29](#)). Certain branches of the tree which appear in the shortest tree, but disappear, or 'collapse' in the tree one step longer, are not drawn in the strict consensus tree. Greater the decay index value, more robust is that internode of the cladogram.

Branches of the tree may also be tested by comparing the number of genetic changes leading up to a particular group, and the CI of individual characters involved. Doyle et al., (1994) on the basis of morphological data, developed a tree having 18 character changes leading to angiosperms. Of these 18 characters 11 had CI of 1, thus

supporting the view that angiosperms form a unique group of plants.

Bootstrap Analysis







Any realistic analysis requires that the data used is randomized. Many techniques are available for randomizing the data. **Bootstrap analysis** is the commonly used method developed by Bradley Efron (1979). Its use in phylogeny estimation was introduced by Felsenstein (1985). Matrix in the Figure 8.28-A contains information on the basis of which the unrooted tree in Figure 8.17-C is constructed. Without touching the rows, any column is chosen at random to become the first column; similarly any other as second and the process is repeated till the number of columns in the new matrix is the same as in the original matrix. As the columns are picked up from the original matrix, the new matrix may contain some characters represented several times (the same column may have been picked up at random more than once), while others may have been omitted (the columns were not picked up at all). The method is known as random sampling with replacement. The resultant matrix B shows that character carpel fusion was picked up twice, whereas the random selection process missed the stamen number.

Repeating the method of random selection, multiple such matrices (usually more than 100) are constructed, and for each matrix the most parsimonious tree/trees found. The consensus tree is developed from these most parsimonious trees. In this consensus tree, the percentage number of trees (generated by bootstrap analysis) that contain that clade is indicated as **bootstrap support** value of that clade. Bootstrap analysis based on the assumption that differential weighting by resampling of the original data will tend to produce same clades if the data are good, and reflect actual phylogeny and very little of homoplasy. A bootstrap value of 70 per cent or more is generally considered as good support to the clade.

Several variations of bootstrap analysis are available. **The partial bootstrapping** in-

volves sampling fewer than the full number of characters. The user is asked for the fraction of characters to be sampled. **Block-bootstrapping** is useful for handling correlated characters. When this is thought to have occurred, we can correct for it by sampling, not individual characters, but blocks of adjacent characters. Block bootstrap was introduced by Künsch (1989). If the correlations are believed to extend over some number of characters, you choose a block size, B , that is larger than this, and choose N/B blocks of size B . In its implementation here the block bootstrap “wraps around” at the end of the characters (so that if a block starts in the last $B-1$ characters, it continues by wrapping around to the first character after it reaches the last character). Note also that if you have a DNA sequence data set of an exon of a coding region, you can ensure that equal numbers of first, second, and third coding positions are sampled by using the block bootstrap with $B = 3$. **Partial block-bootstrapping** is similar to partial bootstrapping except sampling blocks rather than single characters.

Jackknife analysis (Jackknifing) is similar to bootstrap analysis but differs in that each randomly selected character may be resampled only once, and not multiple times, and the resultant resampled data matrix is smaller than the original. **Delete-half-jackknifing** involves sampling a random half of the characters, and including them in the data but dropping the others. The resulting data sets are half the size of the original, and no characters are duplicated. The random variation from doing this should be very similar to that obtained from the bootstrap. The method is advocated by Wu (1986). **Delete-fraction jackknifing** was advocated by Farris et. al. (1996) and involves deleting a fraction $1/e$ ($1/2.71828$). This retains too many characters and will lead to overconfidence in the resulting groups when there are conflicting characters. This and the preceding options form a part of the **SEQBOOT** program of **Phylip** software, and the user is asked to supply the fraction of

A		Habit	Carpels	Stamens	Petals
	Plants	Herbaceous	United	>2	Yellow
	Plants	Herbaceous	Free	>2	Yellow
	Plants	Herbaceous	Free	>2	Red
	Plants	Woody	Free	>2	Red
	Plants	Woody	United	>2	Red
	Plants	Woody	United	2	Red







B		Petals	Carpels	Habit	Carpels
	Plants	Yellow	United	Herbaceous	United
	Plants	Yellow	Free	Herbaceous	Free
	Plants	Red	Free	Herbaceous	Free
	Plants	Red	Free	Woody	Free
	Plants	Red	United	Woody	United
	Plants	Red	United	Woody	United

Figure 8.28 **A:** Matrix based on the tree 9:16A. **B:** One possible matrix after procedure of random sampling with replacement.

characters that are to be retained. The program also offers **permuting** method, with following alternatives. **Permuting species within characters** involves permuting the columns of the data matrix separately. This produces data matrices that have the same number and kinds of characters but no taxonomic structure. It is used for different purposes than the bootstrap, as it tests not the variation around an estimated tree but the hypothesis that there is no taxonomic structure in the data: if a statistic such as number of steps is significantly smaller in the actual data than it is in replicates that are permuted, then we can argue that there is some taxonomic structure in the data (though perhaps it might be just the presence of a pair of sibling species). **Permuting characters** simply permutes the order of the characters, the same reordering being applied to all species. It is included as a possible step in carrying out a permutation test of homogeneity of characters (such as the Incongruence Length Difference test). **Permuting characters separately for each species** permute data so as to destroy all phylogenetic structure, while keeping the base composition of each species the same as be-

fore. It shuffles the character order separately for each species.

It is a common practice, and consequently more informative, to indicate the branch length (number of steps needed to reach that clade), bootstrap or jackknife support and Bremer support (decay index) for each clade in the consensus tree (Figure 8.29).

Effect of Different Outgroups

An important component of procedures generating rooted trees is the incorporation of an outgroup in the analysis. In morphological data, the outgroup choice can influence phylogenetic inference. In molecular data, one specific concern is the levels of sequence divergence between outgroups and ingroups and the subsequent possibility of spurious long-branch attraction (Albert et al., 1994). The robustness of tree can be tested by using randomly-generated outgroup sequences, excluding all outgroups, and using outgroups selectively. Sytsma and Baum (1996), investigating the molecular phylogenies of angiosperms, found that removal of all outgroups generated 27 shortest unrooted trees. Using *Ginkgo* only as outgroup yielded

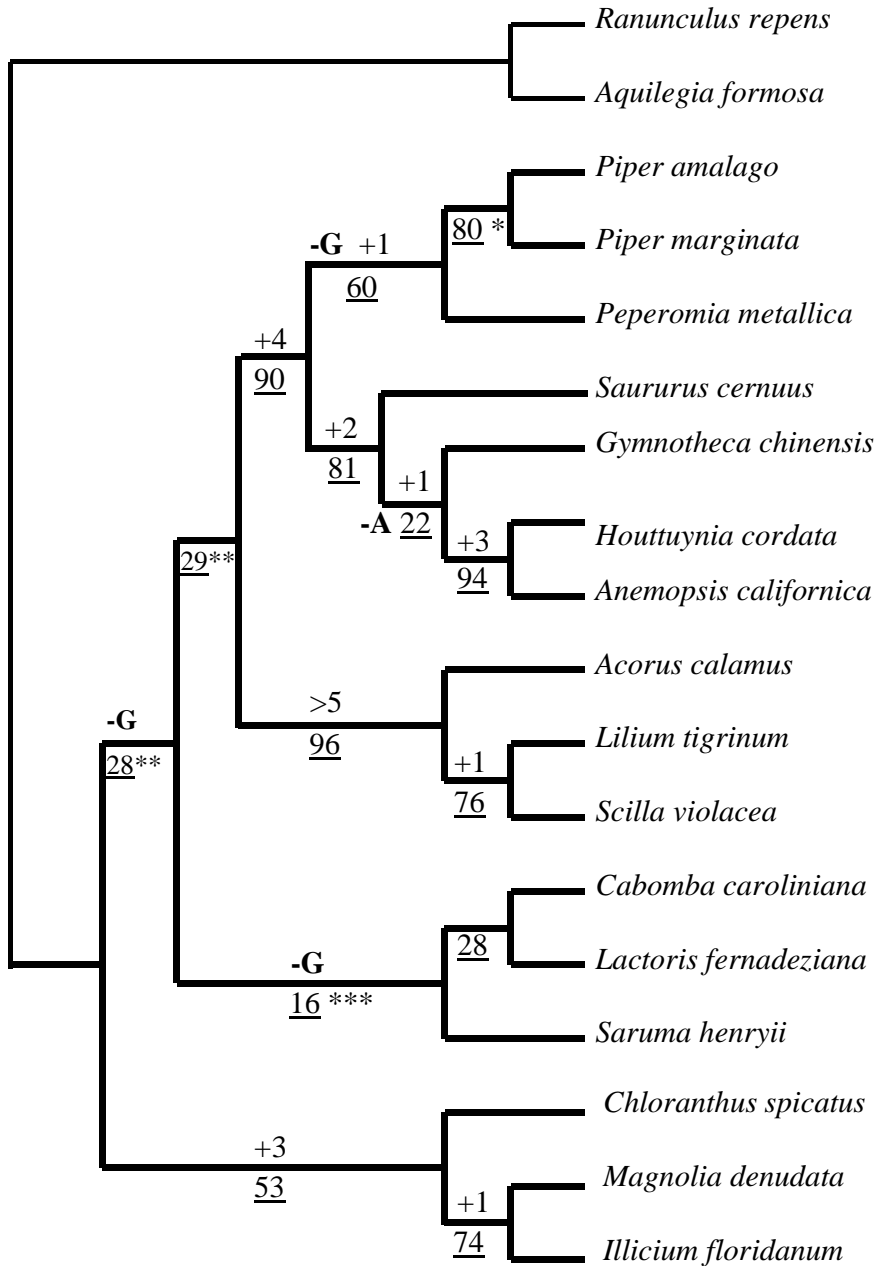


Figure 8.29 Tree developed from the study of 16 species of paleoherbs and 2 outgroup taxa, using 58 morphological and ontogenetic characters. The cladogram requires 214 steps and has CI = 0.51 and RI = 0.65. Bootstrap values are underlined and indicated below a branch. Decay Index is indicated above the branch. *Ranunculus repens* and *Aquilegia formosa* were chosen as outgroup taxa. (Drawn from Tucker and Douglas, 1996).

lineages identical with baseline study (which included all outgroups); when only conifers were used as outgroup, the consensus tree was less resolved and many nodes collapsed. Use of Gnetales as outgroup increased the number of steps needed to yield baseline topologies, and interestingly, *Ceratophyllum* is shown as sister to all angiosperms except eudicots.

Effect of Lineage Removal

Lineage removal strategy highlights the problems of lineage extinction, which often leads to a particular group (especially critical in angiosperms where fossil record is meager) not being sampled in analysis, thus giving distorted phylogenies. The same may also be true for extant taxa, for which very little data is available. The removal of all major lineages, one at a time (Sytsma and Baum, 1994), provided useful information. The removal of *Ceratophyllum*, paleoherbs IIb (Chloranthaceae, and Magnoliales) had no effect on the remaining angiosperm topology, whereas the removal of paleoherbs I (Aristolochiales and Illiciales), Laurales and eudicots showed substantial changes.

Effect of Exemplars

The large computational load in handling a large data is often reduced by using **placeholders** or **exemplars**. These are often used to represent large lineages. The use of exemplars can warn about the possible artifacts when sparsely-sampled lineages appear in basal positions. In such cases, more taxa can be added to the data set for further analyses. But in the case of basal clade where a large number of taxa are extinct, the results could be ambiguous. The results from angiosperms have shown that clades shift around with ease when the number of taxa sampled for each lineage is reduced, and the use of exemplars at times could give misleading results.

Automated Trees

A number sophisticated computer programs are available to construct phylogenetic trees.

These programs are basically similar to those designed for development of phenograms, but differing essentially in the requirement to select one taxon for rooting in most programs. **PHYLIP** (Phylogeny Inference package), is a commonly used set of programs for inferring phylogenies (evolutionary trees) by parsimony, compatibility, distance matrix methods, and likelihood. It can also compute consensus trees, compute distances between trees, draw trees, resample data sets by bootstrapping or jackknifing, edit trees, and compute distance matrices. It can handle data that are nucleotide sequences, protein sequences, gene frequencies, restriction sites, restriction fragments, distances, discrete characters, and continuous characters. Distance matrix can be generated using programs such as **DNADIST** (which handles nucleotide sequence data; it gives you choice to set weightage for transversions/transitions), **PRODIST** (which works with protein sequences) and **RESTDIST** (which works with restriction site data). The most commonly used programs of PHYLIP for handling distance matrix data include **FITCH**, **KITSCH**, and **NEIGHBOR**. These deal with data which comes in the form of a matrix of pairwise distances between all pairs of taxa. **NEIGHBOR** offers **UPGMA** option in which no taxon needs to be selected for rooting, whereas neighbor-joining option of this program, as well as **FITCH** and **KITSCH** need one taxon to be selected for rooting, otherwise by default first taxon is used for rooting. The outtree generated by these programs can be plotted using **DRAWGRAM** or **DRAWTREE**, latter plotting only unrooted trees. **DRAWGRAM** provides a variety of options to choose from. The trees can be drawn horizontal or vertical, branches square (phenogram), v-shaped (cladogram), curved or circular. The branch lengths may be depicted (phylogram) on the tree. The DNA sequence data presented in Figure 8.22 was analysed using PHYLIP programs. Outputs are presented in Figure 8.30. DNA sequence data can also be handled by **DNAPARS** program which performs Parsimony analysis

6 53

Taxona GCCAACGTCG ATGCCAGGTT GTTTAGCACC GGTCTTGTC CGATCACAGA TGT
 Taxonb GCCAACAAATG ATACCAGGCC GTCCAGCACC GATTCTCGTC CGAGTACCGA TGT
 Taxonc GGTAACGTCA ATGGGACGTT GTCCAGCACC GGTTCATGTC CAAGCAGAGA TGT
 Taxond GCCAACATTG ATACCAGGCC GTTTAGCTGC GACTCTGTC CGATCACAA TGT
 Taxone GCTAACGACA ATACCAGGCT GTCCAGCTCC GGTTCAGTC CGAGCACAGA TGT
 Taxonf GCCAACATCG ATGGGACGTT GTTTAGCTCC GATTTCATGTC CAATCACAA TGT

I

6

Taxona 0.000000 0.297888 0.253844 0.301576 0.306645 0.199728
 Taxonb 0.297888 0.000000 0.470560 0.229212 0.276609 0.401604 (((Taxona:0.09986,Taxonf:0.09986):0.03265,Taxonc:0.13251):0.04302,
 Taxonc 0.253844 0.470560 0.000000 0.736034 0.286112 0.276199 Taxone:0.17553):0.02684,(Taxonb:0.11461,Taxond:0.11461):0.08776);
 Taxond 0.301576 0.229212 0.736034 0.000000 0.475079 0.278527
 Taxone 0.306645 0.276609 0.286112 0.475079 0.000000 0.460436
 Taxonf 0.199728 0.401604 0.276199 0.278527 0.460436 0.000000

II

III

IV

Taxono GGCAACGACG ATACCAGGTT GTTTAGCTCC GGTCTCGTC CCAGCAGCCA TGT

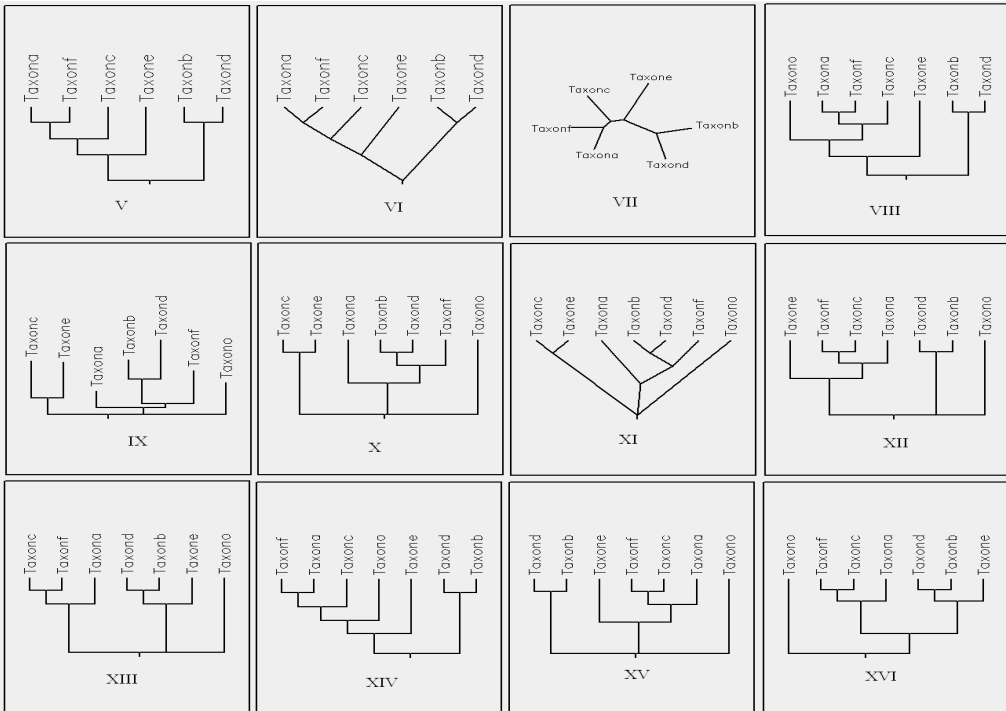


Figure 8.30 Analysis of the DNA sequence data presented in Figure 8.22 using PHYLIP. I: Infile, first line indicating number of taxa and number of nucleotides in each sequence; II: Square distance matrix (outfile) generated by DNADIST program; III: Outtree file generated by NEIGHBOR program using UPGMA option; IV: DNA sequence of the 7th hypothetical taxon (taxono) used for rooting; V-VI: Square tree (Phenogram) and V-shaped tree (cladogram); VII: Unrooted tree of same; VIII-XVI: Diagrams based on 7-taxa sequences; VIII: Phenogram, UPGMA option; IX-XI: Phylogram, Phenogram and Cladogram based on neighbour-joining option of NEIGHBOR; XII: Phenogram based on DNAML program; XIII: Phenogram based on FITCH program; XIV: Phenogram based on KITSCH program; XV: Tree (Phenogram) generated based on DNAPARS program; XVI: Majority-rule consensus tree based on CONSENSE program, using outtree files of above six programs. (All trees except VII (plotted using DRAWTREE) plotted using different options of DRAWGRAM program).

15 9

Taxon1	101001111
Taxon2	101011001
Taxon3	010000111
Taxon4	111010000
Taxon5	101101101
Taxon6	110011010
Taxon7	011011001
Taxon8	000000111
Taxon9	111001100
Taxon10	000010111
Taxon11	101101001
Taxon12	110011001
Taxon13	001010101
Taxon14	101010101
Taxon15	000000000

A

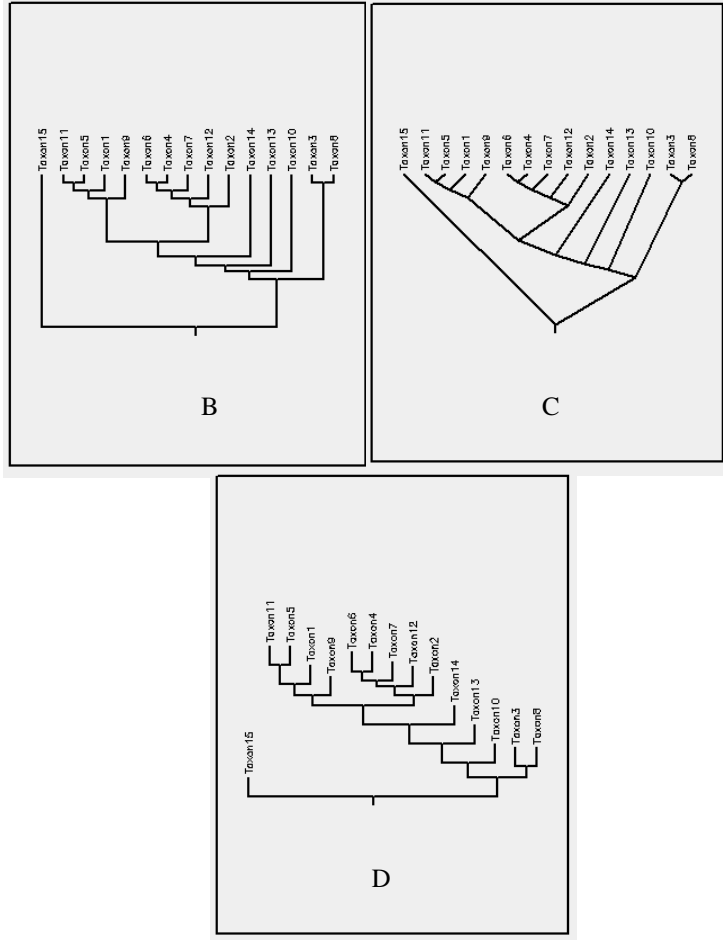


Figure 8.31 Construction of trees using MIX program of PHYLIP based on matrix in the Table 8.4. **A:** Input file with fourth character converted into binary (simple and compound leaves) character. Out of the 34 parsimonious trees generated by Mix, Consensus tree generated by CONSENSE program presented as Phenogram (**B**), Cladogram (**C**) and Phylogram (**D**).

and selects the best tree. It gives you choice to select the number of trees to be saved, 10000 being the default. The program directly yields the outtree file. PROTPARS, similarly performs Parsimony analysis of Protein sequences. The protein sequences are given by the one-letter code used by the late Margaret Dayhoff's group in the Atlas of Protein Sequences, and consistent with the IUB standard abbreviations. DNAMOVE which handles data similar to DNAPARS,

allows the user to choose an initial tree, and displays this tree on the screen. The user can look at different sites and the way the nucleotide states are distributed on that tree, given the most parsimonious reconstruction of state changes for that particular tree. The user then can specify how the tree is to be rearranged, rerooted or written out to a file. By looking at different rearrangements of the tree the user can manually search for the most parsimonious tree, and

can get a feel for how different sites are affected by changes in the tree topology. DNAML program carries out analysis of DNA sequences using Maximum Likelihood Method. The program uses both informative and non-informative sites and yields the outtree file directly. RESTML similarly handles restriction site data using maximum likelihood method. Binary data coded as 0 (ancestral state) and 1 (advanced state) is handled by MIX, which performs parsimony analysis and generates outtree which can be plotted using DRAWGRAM. Input data from Table 8.4 and most parsimonious tree generated using MIX program is presented in Figure 8.31. For this analysis fourth multistate character was converted into binary character (simple and compound leaves). Using Wagner parsimony the program was able to generate 34 trees. Taxon 15 was used for rooting. CONSENSE program was used to select the majority rule consensus tree. MOVE handles binary data and is an interactive program which allows the user to choose an initial tree, and displays this tree on the screen. The user can look at different characters and the way their states are distributed on that tree, given the most parsimonious reconstruction of state changes for that particular tree. The user then can specify how the tree is to be rearranged, rerooted or written out to a file. By looking at different rearrangements of the tree the user can manually search for the most parsimonious tree, and can get a feel for how different characters are affected by changes in the tree topology.

Multistate data can similarly be handled by PARS, and can be converted into binary data by FACTOR program. Data from Gene frequencies and continuous characters is handled by CONTML (constructs maximum likelihood estimates of the phylogeny; handles both types of data), GENDIST (computes genetic distances for use in the distance matrix programs; handles data from gene frequencies) and CONTRAST (examines correlation of traits as they evolve along a given phylogeny; handles continuous characters data). The data matrix for gene fre-

quencies contains number of species (or populations) and number of loci, where as the second line contains number of alleles for each locus. the default number of data for each species (A-all) contains one allele less for each locus. thus for three loci with 2, 3 and 2 alleles respectively there would be four values. Without A option, there should be 7 values. The values in dataset are preceded and followed by blanks. The data from continuous characters does not contain the second line, the data would include number of species and the number of characters in the first line (only line above species data).

PHYLIP also offers programs to yield consensus tree (CONSENSE), Bootstrapping (SEQBOOT) and a host of related programs. The following information may be useful in handling DNA sequence data.

Prepare infile of DNA sequences in which taxon name takes 10 characters followed by sequences in groups of 10 (separated by a space), last three nucleotides being terminating codon. First taxon should be one intended as one used for rooting. Number of taxa (sequences used) and number of nucleotides in each sequence forms first line of file. Longer sequences can be interleaved (giving first part of sequences of all taxa and then next part of all taxa) or aligned (finishing one sequence and then going to second). Save this file in text format in notepad (ANSI code should be used; is default in notepad). Distance matrix can be prepared using dnadist.exe program. When program asks for infile, type above file name along with .txt ending. Choose the ratio of transitions and transversions, so that program can handle it accordingly. You can also choose distance model such as F84, Kimura, Jukes-Cantor and Logdet. Give name of your output file (preferably in txt format so that you can open and see it in notepad). The file can be saved as square matrix or only lower triangle. The above file can be used for generating clusters through Fitch, Kitsch, and Neighbor programs. Each program searches for the shortest tree. When program asks for infile, type the name of above output file. Give name of

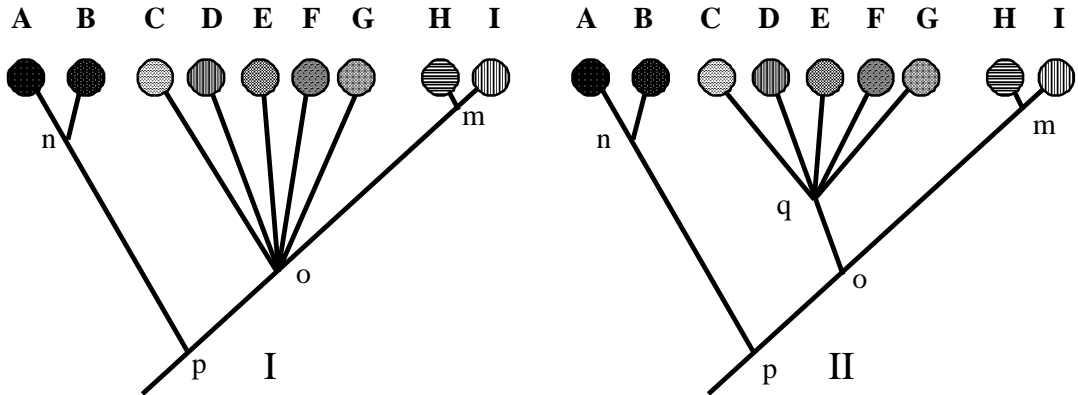


Figure 8.32 Attempts towards construction of monophyletic groups. **I:** Strict consensus tree as presented in the Figure 8.27-III. With poorly resolved phylogenies, the separation of H and I in a group distinct and of the same rank as group CDEFG would create a paraphyly, as HI are left out of the descendents of common ancestor **o**. **II:** A consensus tree (hypothetical) with better resolved phylogenies. Both groups CDEFG and HI are monophyletic and, in turn could be assembled into more inclusive group with common ancestor at level **o**, now containing all descendents of the common ancestor. This group (CDEFG, HI) and AB (also monophyletic) could be assembled into one most inclusive monophyletic group, containing all descendents of the common ancestor at **p**.

output file (of this program) or simply ask for replacement if program reports that file is already present. Neighbour provides a choice between Neighbour-joining (in which one taxon is to be chosen for rooting) and UPGMA (in which no taxon for rooting has been selected). After selection of choice press Y. Program will generate outtree file, if already present replace it. You can read this file if saved in txt format. The above outtree file can be used for plotting trees using DRAWGRAM or DRAWTREE programs. Draw program asks for intree file. Type in the name of above outtree file. It will next ask for name of font file. Type font1 or any other within the Phylip folder. The program provides you number of choices including phenogram and cladogram. It also provides choice between indicating branch lengths (construction of conventional phylogram) or not (conventional phenograms and cladograms where taxa end at same height). On typing Y tree preview will appear. Press Print screen on keyboard and paste on new

paint file to save as image file in Paint. You can change options by clicking File->change parameters in tree preview and go back to drawgram to generate other types of trees.

Binary data can similarly be input in infile with just replacing nucleotide alphabets with binary 0 and 1 data as presented in Table 8.31 and handled by various programs mentioned earlier.

Gene Trees and Species Trees

Traditionally the phylogenetic trees are constructed using data from multiple characters, and if genetic data is used, from analysis of multiple genes. Such trees, appropriately known as **species trees** reflect the evolutionary history of related groups of species, and consequently a single species. A phylogenetic tree based on the divergence observed within a single homologous gene is most appropriately called a **gene tree**. The genes commonly used for the construction of gene trees have been described in

chapter 7. Although they have been broadly used in recent years in the construction of phylogenies, a single gene may not always reflect relationships between species, because divergence within genes, especially the sequence polymorphism occurs before the splitting of populations that give rise to new species.

Developing Classification

Once the phylogeny of a group has been developed, the evolutionary process within the group can be reconstructed, the morphological, physiological and genetic changes can be described, and the resultant information used in the classification of the group. Phylogenetic classifications are based on the recognition of monophyletic groups and avoid including paraphyletic and often completely reject paraphyletic groups. Such classifications are superior over classifications based on overall similarity in several respects:

1. Such a classification reflects the genealogical history of the group much more accurately.
2. The classification based on monophyletic groups is more predictive and of greater value than classification based on some characteristics.
3. Phylogenetic classification is of major help in understanding distribution patterns, plant interactions, pollen biology, dispersal of seeds and fruits.
4. The classification can direct the search for genes, biocontrol agents and potential crop species.
5. The classification can be of considerable help in conservation strategies.

The evolutionary history of the of the group of 8 living species shown in Figure 8.12 was known with precise point of character transformations, and the construction of monophyletic groups, assembled into successively more inclusive groups, did not pose much problem. But it is often not the case. Even, most resolved consensus trees are often ambiguous in several respects.

Consider the strict consensus tree represented in the Figure 8.27-III. This tree is reproduced in the Figure 8.32-I. As noted earlier, the phylogenetic relationships between taxa (these could be different species, genera, etc.) C, D, E, F, and G are poorly resolved, and as such they are shown arising from the common point, and consequently common ancestor as level **o**. Although H and I form a distinct group with a common ancestor as **m**, but leaving these two out of the group including CDEFG would render latter as paraphyletic (cf. traditional separation of dicots and monocots). The safest situation would be to include all the seven taxa into one group, which may be regarded as belonging to the same rank as the group including A and B. All the nine taxa may next be included into the single most inclusive group with common ancestry at level **p**. We are thus able to construct groups at two ranks only.

Now supposing the phylogenies of the taxa were better resolved and we had obtained a consensus tree as shown in Figure 8.32-II. Now taxa C,D,E,F and G belong to a lineage which diverged from the main lineage, successive to the divergence of the lineage formed by A and B. Placement of H and I into one group HI would not create any problem as both this group as well as the group CDEFG are monophyletic with separate common ancestors at level **m** and **q**, respectively. The groups CEDFG and HI could next be assembled into group CDEFGHI with common ancestor at **o**. Note that the group AB can next be merged with CDEFGHI to form single most inclusive group ABCDEFGHI. Now we have been able to construct taxa at three ranks instead of two from tree I.

Supposing the taxa A to I included in the tree, are different species. From tree II, thus we are able to recognize three genera AB, CDEFG and HI. The last two are next assembled into family CDEFGHI and the former a monotypic family AB. The other alternative was to place A and B in two separate monotypic genera (depending on the degree of morphological and genetic divergence obtained) which are then assembled into

family AB. The two families may next be assembled into order ABEDEFGHI. There could be other possibilities also. The second rank could be a subfamily and the third a family. Similarly, a third rank could be a suborder instead of an order. These final decisions are often made, based on the size of the group, degree of divergence, and the reliability of characters. All the groups recognized above would be monophyletic at the respective ranks.

Next, let us look at the tree shown in the Figure 8.29, a study on paleoherbs. *Ranunculus* and *Aquilegia* were used as outgroup representing family Ranunculaceae; their isolated position from paleoherbs is clearly depicted in the tree. Paleoherbs constitute a group of taxa of uncertain affinities, which have been placed differently in various classification schemes, but a few points seem to have been resolved. *Piper*, and *Peperomia* (both belong to Piperaceae) form a distinct group, and so do *Saururus*, *Gymnotheca*, *Houttuynia* and *Anemopsis* (all four belonging to Saururaceae), and the two families a well supported (bootstrap support of 90 per cent). This was confirmed by comparison of seven published trees of paleoherbs. *Cabomba*, *Lactoris* and *Saruma* have least resolved

affinities with very poor support, with highly unstable position.

The final decisions on the recognition of groups are, however, often based on personal interpretation of phylogenies. *Chloranthus*, in this tree as well in several others, is closer to *Magnolia* (Magnoliales) and *Laurus* (Laurales), but often finds different treatment. APG II places Chloranthaceae after Amborellaceae at the start of Angiosperms. Judd et al., had earlier (1999) placed Chloranthaceae under order Laurales of Magnoliid complex, but have now (2008) shifted the family among basal ANITA Grade with uncertain position. APweb of Stevens (2008), which places the family under order Chloranthales. Thorne had earlier (1999, 2000) placed Chloranthaceae in Magnoliidae → Magnoliales → Chloranthaceae (other suborders within the order being Magnoliaceae, and Lauraceae), but subsequently (2003) included the family after Amborellaceae under order Chloranthales, the first order of Magnoliidae, finally (2006, 2007) separated under subclass Chloranthidae, a placement somewhat similar to APG II. Further discussion on angiosperm affinities will be resumed in the next chapter.