Cladistics & Phylogenetics

Frida Brill

Joshua Krieger

Revised Edition: 2014

ISBN 978-81-323-0689-4

Published by:
Academic Studio
4735/22 Prakashdeep Bldg,
Ansari Road, Darya Ganj,
Delhi - 110002
Email: info@wtbooks.com

Table of Contents

Chapter- 1

Introduction to Cladistics

Cladistics is a method of classifying species of organisms into groups called **clades**, which consist only of firstly, all the descendants of an ancestral organism and secondly, the ancestor itself. For example, birds, dinosaurs, crocodiles, and all descendants (living or extinct) of their most recent common ancestor form a clade. In the terms of biological systematics, a clade is a single "branch" on the "tree of life", a monophyletic group.

Cladistics can be distinguished from other taxonomic systems, such as phenetics, by its focus on shared derived characters (synapomorphies). Systems developed earlier usually employed overall morphological similarity to group species into genera, families and other higher level groups (taxa); cladistic classifications (usually in the form of trees called cladograms) are intended to reflect the relative recency of common ancestry or the sharing of homologous features. Cladistics is also distinguished by an emphasis on parsimony and hypothesis testing (particularly falsificationism), leading to a claim that cladistics is more objective than systems which rely on subjective judgements of relationship based on similarity.

Cladistics originated in the work of the German entomologist Willi Hennig, who referred to it as "phylogenetic systematics" (also the name of his 1966 book); the use of the terms "cladistics" and "clade" was popularized by other researchers. The technique and sometimes the name have been successfully applied in other disciplines: for example, to determine the relationships between the surviving manuscripts of the *Canterbury Tales*.

Cladists use *cladograms*, diagrams which show ancestral relations between species, to represent the monophyletic relationships of species, termed sister-group relationships. This is interpreted as representing phylogeny, or evolutionary relationships. Although traditionally such cladograms were generated largely on the basis of morphological characters, genetic sequencing data and computational phylogenetics are now very commonly used in the generation of cladograms.

Cladistics, either generally or in specific applications, has been criticized from its beginnings. A decision as to whether a particular character is a synapomorphy or not may be challenged as involving subjective judgements, raising the issue of whether cladistics as actually practised is as objective as has been claimed. Formal classifications based on cladistic reasoning are said to emphasize ancestry at the expense of descriptive

characteristics, and thus ignore biologically sensible, clearly defined groups which do not fall into clades (e.g. reptiles as traditionally defined or prokaryotes).

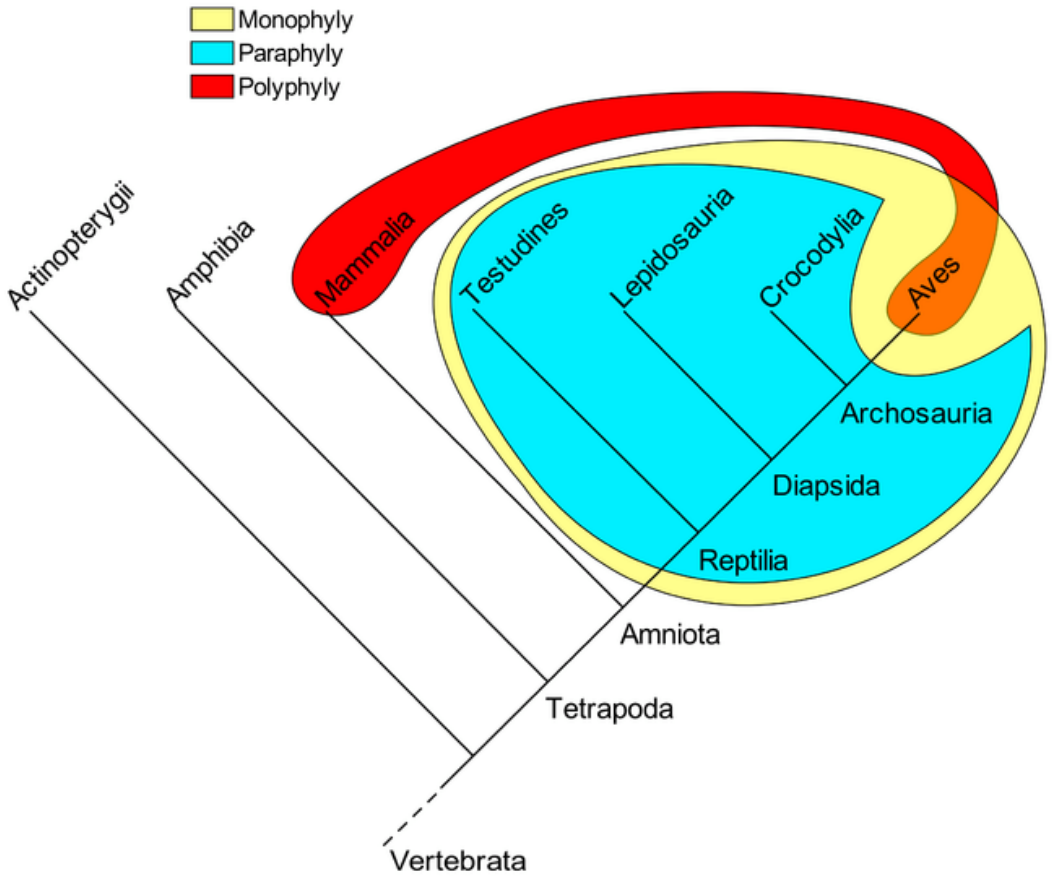
History of cladistics

The term *clade* was introduced in 1958 by Julian Huxley, *cladistic* by Cain and Harrison in 1960, and *cladist* (for an adherent of Hennig's school) by Mayr in 1965. Hennig referred to his own approach as *phylogenetic systematics*. From the time of his original formulation until the end of the 1980s cladistics remained a minority approach to classification. However in the 1990s it rapidly became the dominant method of classification in evolutionary biology. Computers made it possible to process large quantities of data about organisms and their characteristics. At about the same time the development of effective polymerase chain reaction techniques made it possible to apply cladistic methods of analysis to biochemical and molecular genetic features of organisms as well as to anatomical ones.

Cladistics as a successor to phenetics

For some decades in the mid to late twentieth century, a commonly used methodology was phenetics ("numerical taxonomy"). This can be seen as a predecessor to some methods of today's cladistics (namely distance matrix methods such as neighbor-joining), but made no attempt to resolve phylogeny, only similarities.

Clades



Partial Evolutionary Tree of the Vertebrates

The yellow group (sauropsids) is monophyletic, the blue group (traditional reptiles) is paraphyletic, and the red group (warm-blooded animals) is polyphyletic.

A clade is a group of taxa consisting only of an ancestor taxon and all of its descendant taxa. In the diagram provided (a **cladogram**), it is hypothesized that all vertebrates, including ray-finned fishes (Actinopterygii), had a common ancestor, and so form a clade. Within the vertebrates, all tetrapods, including amphibians, mammals, reptiles (as traditionally defined) and birds, are hypothesized to have had a common ancestor, and so also form a clade. The tetrapod ancestor was a descendant of the original vertebrate ancestor, but is not an ancestor of any ray-finned fish living today.

An important caution is that any cladogram is a provisional hypothesis. Although unlikely, future genetic or morphological evidence might suggest that ray-finned fish and amphibians share a common ancestor that was not an ancestor of the other tetrapods. The

new information would cause us to define a ray-finned-fish-and-amphibian clade, altering the cladogram.

The relationship between clades can be described in several ways:

- A clade is *basal* to another clade if it contains that other clade as a subset within it. In the example, the vertebrate clade is basal to the tetrapod and ray-finned fish clades. (Some authors have used "basal" differently to mean a clade that is less species-rich than a sister clade, with such a deficit being taken as an indication of 'primitiveness'. Others consider this usage to be incorrect.)
- A clade located within a clade is said to be *nested* within that clade. In the diagram, the tetrapod clade is nested within the vertebrate clade.
- Two clades are *sisters* if they have an immediate common ancestor. In the diagram, crocodiles and birds are sister clades, as are amphibians and amniotes.

Terminology for characters

The following terms are used to identify shared or distinct characters among groups:

- *Plesiomorphy* ("close form") or *ancestral state*, also *symplesiomorphy* ("shared plesiomorphy", i.e. "shared close form"), is a characteristic that is present at the base of a tree (cladogram). Since a plesiomorphy that is inherited from the common ancestor may appear anywhere in a tree, its presence provides no evidence of relationships within the tree. The traditional definition of reptiles (the blue group in the diagram) includes being cold-blooded (i.e. not maintaining a constant high body temperature), whereas birds are warm-blooded. Since cold-bloodedness is a plesiomorphy, inherited from the common ancestor of traditional reptiles and birds, it should not be used to define a group in a system based on cladistics.
- *Apomorphy* ("separate form") or *derived state* is a characteristic believed to have evolved within the tree. It can thus be used to separate one group in the tree from the rest. Within the group which shares the apomorphy it is a *synapomorphy* ("shared apomorphy", i.e. "shared separate form"). For example, within the vertebrates, all tetrapods (and only tetrapods) have four limbs; thus, having four limbs is an synapomorphy for tetrapods. All the tetrapods can legitimately be grouped together because they have four limbs.
- *Homoplasy* is a characteristic shared by members of a tree but not present in their common ancestor. It arises by convergence or reversion. Both mammals and birds are able to maintain a high constant body temperature (i.e. they are 'warm-blooded'). However, the ancestors of each group did not share this character, so it must have evolved independently. Mammals and birds should not be grouped together on the basis that they are warm-blooded.

The terms (sym)plesiomorphy and (syn)apomorphy are relative and their application depends on the position of a group within a tree. An apomorphy of one clade is a plesiomorphy of another contained within it. For example, when trying to decide whether

tetrapods should form a clade, an important question is whether having four limbs is a synapomorphy of all the taxa to be included within Tetrapoda: did all the possible members of the Tetrapoda inherit four limbs from a common ancestor, whereas all other vertebrates did not? By contrast, for a group within the tetrapods, such as birds, having four limbs is a plesiomorphy. The fact that ostriches and rheas both have four limbs does not provide any support for putting them into a separate group of 'flightless birds'. Using these two terms allows a greater precision in the discussion of homology, in particular allowing clear expression of the hierarchical relationships among different homologies.

It can be difficult to decide whether a character is in fact the same, and thus can classified as a synapomorphy which may identify a group, or whether it only appears to be the same, and is thus a homoplasy which cannot identify a group. There is a danger of circular reasoning: assumptions about the shape of a phylogenetic tree are used to justify decisions about characters, which are then used as evidence for the shape of the tree. It has been argued that this kind of reasoning has been used by proponents of the view that birds are nested within the theropod dinosaur clade.

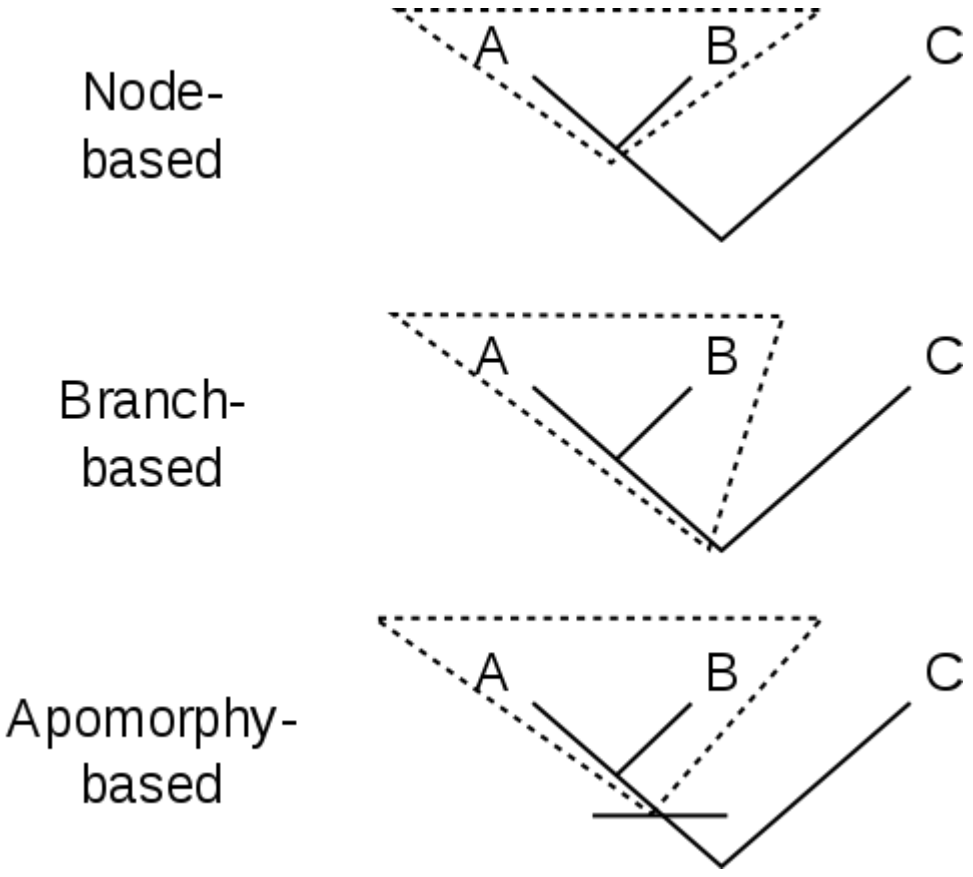
Terminology for groups

Three main types of group can be identified on the basis of their relationships in cladograms. The three can be defined in two different but related ways, as shown in the table below. The first is in terms of the shape of a set of nodes taken from a cladogram. In this approach, an 'ancestor node' is simply a branching point in the diagram; it may or may not correspond to an actual ancestor. The second is in terms of the characters of the taxa being classified and how these characters have been inherited. In this approach, an ancestor is an actual taxon, whether currently known or not.

Term	Node-based definition	Character-based definition
Monophyly	A monophyletic group of nodes in a tree is one which includes all the nodes descended from their most recent common ancestor node, plus the most recent common ancestor node, but no other nodes.	A monophyletic group of taxa is characterized by one or more **synapomorphies**: derived characters inherited by all members of the group from ancestors and not inherited by any other taxa. A monophyletic group is a 'clade'. A 'crown group' is an example of a monophyletic group.
Paraphyly	A paraphyletic group of nodes in a tree is one which is constructed by taking a monophyletic group and removing one or more smaller monophyletic groups. (Removing one group produces a singly	A paraphyletic group of taxa is characterized by one or more **(sym)plesiomorphies**: characters inherited from ancestors but not present in all of their descendants. As a consequence, a paraphyletic group is truncated, in that it excludes one or more monophyletic taxa from

	paraphyletic group, removing two a doubly paraphylectic group, and so on.) A paraphyletic group is necessarily non-monophyletic.	an initially monophyletic group. An alternative name is an 'evolutionary grade', refering to the ancestral character state within the group. A 'stem group' is an example of a paraphyletic group.
Polyphyly	A polyphyletic group of nodes in a tree is one which is neither monophyletic nor paraphyletic.	A polyphyletic group of taxa is characterized by by one or more **homoplasies**: characters which have converged or reverted so as to appear to be the same but which have not been inherited from common ancestors. As a consequence, polyphyletic groups of taxa are totally artificial.

Branch-based definitions of clade



Three alternative ways to define a clade

The node-based definition of a monophyletic group (i.e. a clade) given above regards the lines in the cladogram only as a way of showing connections between taxa. This is appropriate when considering only living (extant) taxa; however, when extinct taxa are to be included in a cladogram, lines correspond to sequences of ancestors. There are two alternative ways of defining a clade which explicitly take into account the line below the branching point at the base of a clade. These definitions are most notably set out in the PhyloCode.

Consider how a clade combining A and B in the diagram can be defined.

- *Node-based*: The node-based definition specifies A+B as the *last* common ancestor of A and B, and all descendants of that ancestor. It thus excludes from the clade the line below the junction of A and B. Crown groups are a type of node-based clade.
- *Branch-based*: A branch-based definition specifies A+B as the *first* ancestor of A which is not also an ancestor of C, and all descendants of that ancestor. It thus includes in the clade the line below the junction of A and B. (This type of definition was originally called "stem-based", but this was changed to avoid confusion with the term "stem group", which is parapyletic.) Total groups are a type of branch-based clade.
- *Apomorphy-based*: An apomorphy-based definition specifies A+B as the first ancestor of A to possess derived trait M homologously (that is, synapomorphically) with that trait in A, and all descendants of that ancestor. It thus includes in the clade only that part of the line below the junction of A and B which corresponds to ancestors possessing the apomorphy. The process of identifying and naming groups based on apomorphies is the method that most resembles classical systematics, with the proviso that cladistic taxa always denote a clade.

Note that these alternative definitions do not alter the classification of the tips of the tree, and so are equivalent if only living (extant) taxa are being considered.

Cladograms

Cladists use *cladograms*, diagrams which show ancestral relations between taxa, to represent the evolutionary tree of life. Although traditionally such cladograms were generated largely on the basis of morphological characters, molecular sequencing data and computational phylogenetics are now very commonly used in the generation of cladograms.

The starting point of cladistic analysis is a group of species and molecular, morphological, or other data characterizing those species. The end result is a tree-like relationship diagram called a cladogram, or sometimes a *dendrogram* (Greek for "tree drawing"). The cladogram graphically represents a hypothetical evolutionary process. Cladograms are subject to revision as additional data become available.

The terms "evolutionary tree", and sometimes "phylogenetic tree" are often used synonymously with cladogram but others treat phylogenetic tree as a broader term that includes trees generated with a nonevolutionary emphasis. In cladograms, all species lie at the leaves. The two taxa on either side of a split, with a common ancestor and no additional descendents, are called "sister taxa" or "sister groups". Each subtree, whether it contains only two or a hundred thousand items, is called a "clade". Many cladists require that all forks in a cladogram be 2-way forks. Some cladograms include 3-way or 4-way forks when there are insufficient data to resolve the forking to a higher level of detail.

For a given set of taxa, the number of distinct cladograms that can be drawn (ignoring which cladogram best matches the taxon characteristics) is:

Number of taxa	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**N**
Number of rooted cladograms	1	3	15	105	945	10,395	135,135	2,027,025	34,459,425	1*3*5*7*...*(2N-3)

This superexponential growth of the number of possible cladograms explains why manual creation of cladograms becomes very difficult when the number of taxa is large. If a cladogram represents N taxa, the number of levels (the "depth") in the cladogram is on the order of $\log_2(N)$. For example, if there are 32 species of deer, a cladogram representing deer could be around 5 levels deep (because $2^5 = 32$), although this is really just the lower limit. A cladogram representing the complete tree of life, with about 10 million species, could be about 23 levels deep. This formula gives a lower limit, with the actual depth generally a larger value, because the various branches of the cladogram will not be uniformly deep. Conversely, the depth may be shallower if forks larger than 2-way forks are permitted.

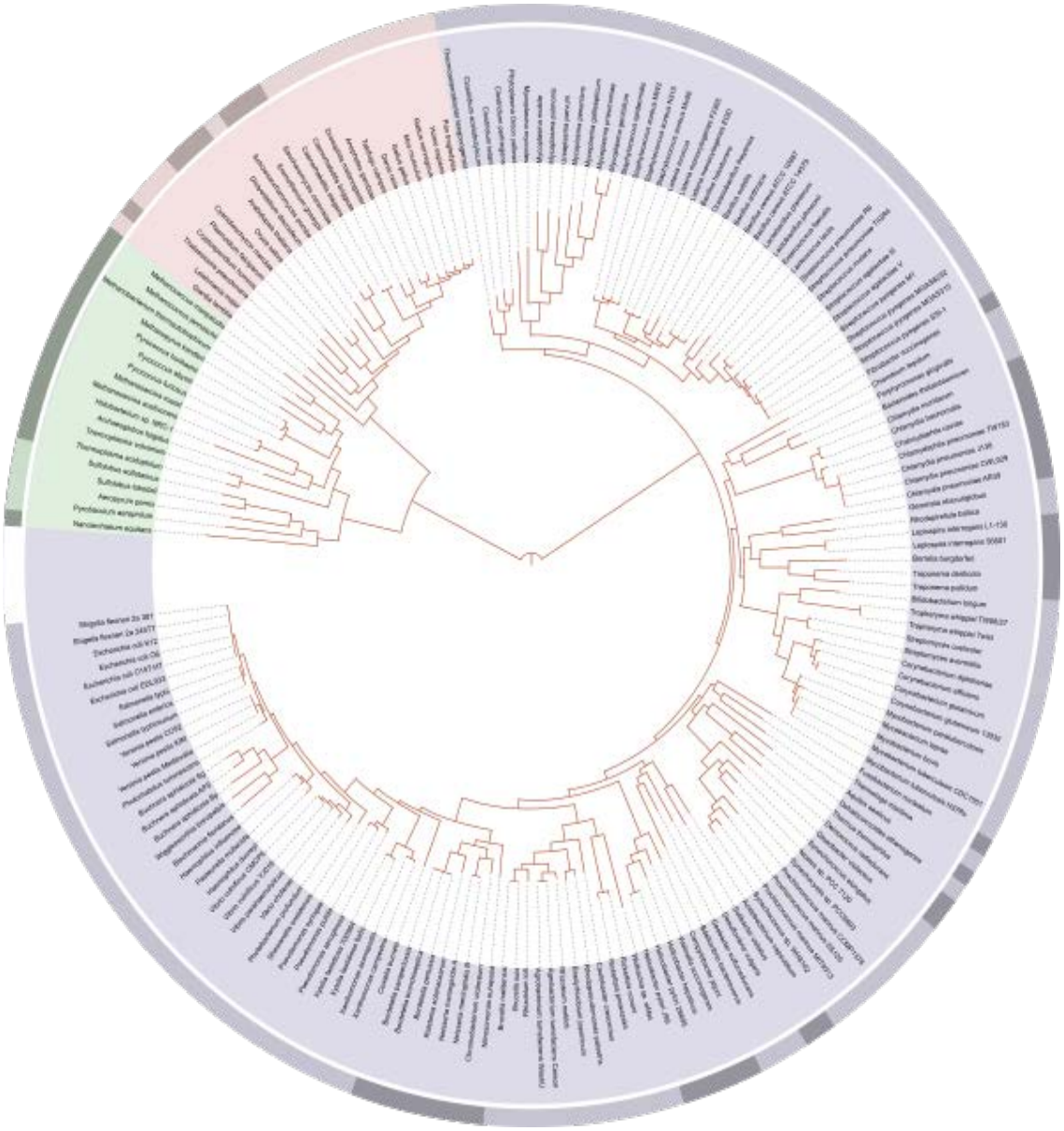
A cladogram tree has an implicit time axis, with time running forward from the base of the tree to the leaves of the tree. If the approximate date (for example, expressed as millions of years ago) of all the evolutionary forks were known, those dates could be captured in the cladogram. Thus, the time axis of the cladogram could be assigned a time scale (e.g. 1 cm = 1 million years), and the forks of the tree could be graphically located along the time axis. Such cladograms are called *scaled cladograms*. Many cladograms are not of this type, for a variety of reasons:

- They are built from species characteristics that cannot be readily dated (e.g. morphological data in the absence of fossils or other dating information)
- When the characteristic data are DNA/RNA sequences, it is feasible to use sequence differences to establish the relative ages of the forks, but converting those ages into actual years requires a significant approximation of the rate of change
- Even when the dating information is available, positioning the cladogram's forks along the time axis in proportion to their dates may cause the cladogram to become difficult to understand or hard to fit within a human-readable format

Cladistics makes no distinction between extinct and extant species, and it is appropriate to include extinct species in the group of organisms being analyzed. Cladograms that are based on DNA/RNA generally do not include extinct species because DNA/RNA samples from extinct species are rare. Cladograms based on morphology, especially morphological characteristics that are preserved in fossils, are more likely to include extinct species.

Cladistics in taxonomy

Phylogenetic nomenclature contrasted with traditional taxonomy



A highly resolved, automatically generated tree of life based on completely sequenced genomes

Most taxonomists have used the traditional approaches of Linnaean taxonomy and later Evolutionary taxonomy to organize life forms. These approaches use several fixed levels of a hierarchy, such as kingdom, phylum, class, order, and family. Phylogenetic nomenclature does not feature those terms, because the evolutionary tree is so deep and so complex that it is inadvisable to set a fixed number of levels.

Evolutionary taxonomy insists that groups reflect phylogenies. In contrast, Linnaean taxonomy allows both monophyletic and paraphyletic groups as taxa. Since the early 20th century, Linnaean taxonomists have generally attempted to make at least family- and lower-level taxa (i.e. those regulated by the codes of nomenclature) monophyletic. Ernst Mayr in 1985 drew a distinction between the terms cladistics and phylogeny: "It would seem to me to be quite evident that the two concepts of phylogeny (and their role in the construction of classifications) are sufficiently different to require terminological distinction. The term *phylogeny* should be retained for the broad concept of phylogeny, promoted by Darwin and adopted by most students of phylogeny in the ensuing 90 years. The concept of phylogeny as mere genealogy should be terminologically distinguished as *cladistics*. To lump the two concepts together terminologically could not help but produce harmful equivocation."

Willi Hennig's pioneering work provoked a spirited debate about the relative merits of phylogenetic nomenclature versus Linnaean or evolutionary taxonomy, which has continued down to the present; however Hennig did not advocate abandoning the Linnaean nomenclatural system. Some of the debates in which the cladists were engaged had been running since the 19th century, but they were renewed fervor, as can be seen from the *Foreword* to Hennig (1979) by Rosen, Nelson, and Patterson:

"Encumbered with vague and slippery ideas about adaptation, fitness, biological species and natural selection, neo-Darwinism (summed up in the "evolutionary" systematics of Mayr and Simpson) not only lacked a definable investigatory method, but came to depend, both for evolutionary interpretation and classification, on consensus or authority."

Phylogenetic nomenclature strictly and exclusively follows phylogeny and has arbitrarily deep trees with binary branching: each taxon corresponds to a clade. Linnaean taxonomy, while since the advent of evolutionary theory following phylogeny, also may subjectively consider similarity and has a fixed hierarchy of taxonomic ranks, and its taxa are not required to correspond to clades.

Paraphyletic groups discouraged

Many cladists discourage the use of paraphyletic groups in classification of organisms, because they detract from cladistics' emphasis on clades (monophyletic groups). In contrast, proponents of the use of paraphyletic groups argue that any dividing line in a cladogram creates both a monophyletic section above and a paraphyletic section below. They also contend that paraphyletic taxa are necessary for classifying earlier sections of the tree – for instance, the early vertebrates that would someday evolve into the family