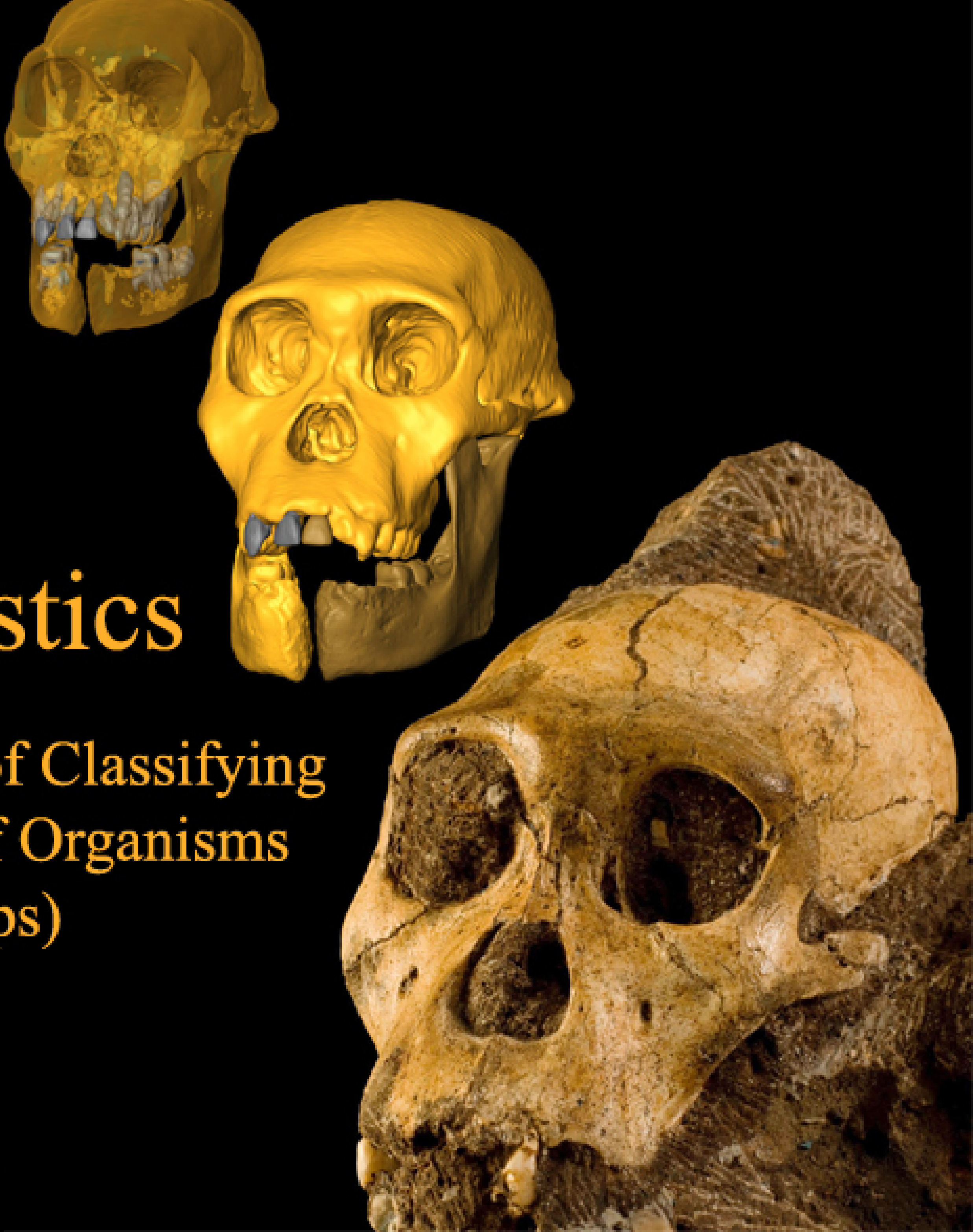


Cladistics

(Method of Classifying Species of Organisms into Groups)

Frida Brill

First Edition, 2012

ISBN 978-81-323-3726-3

Published by:
University Publications
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).

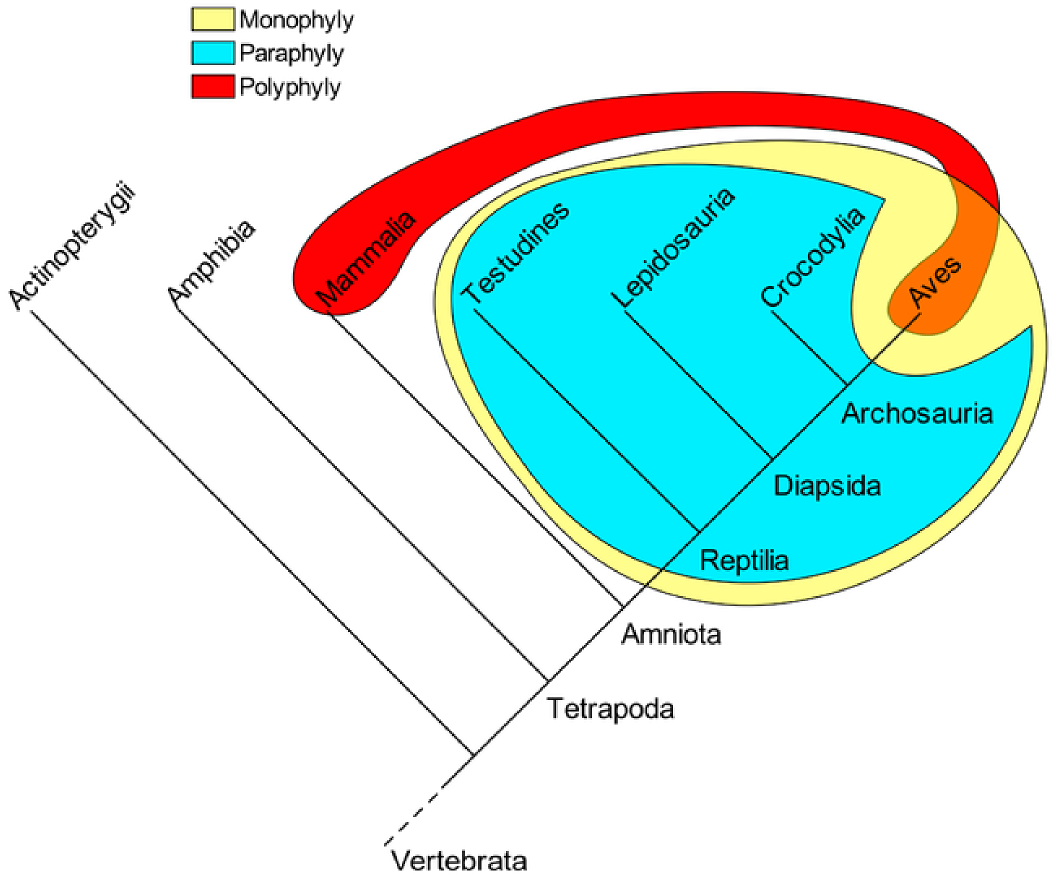
History of cladistics

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.

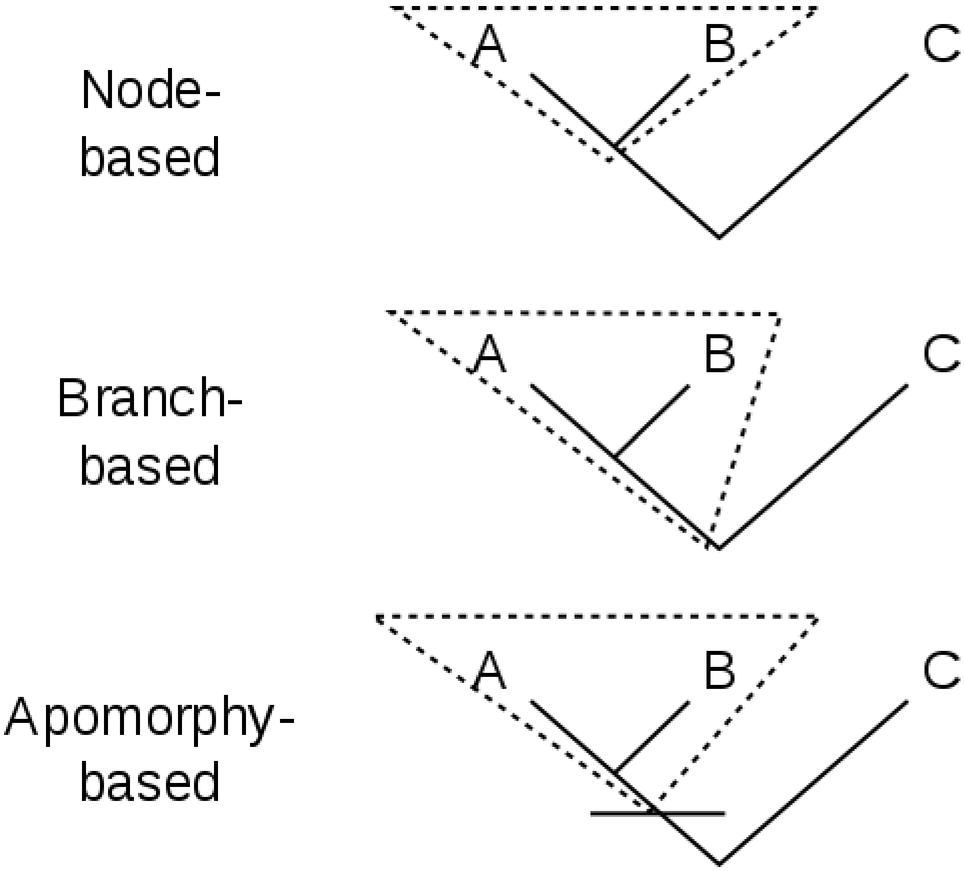
Terminology for groups

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 paraphyletic group, removing two a doubly paraphylectic group, and so on.) A paraphyletic group is necessarily non-	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 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

	monophyletic.	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.

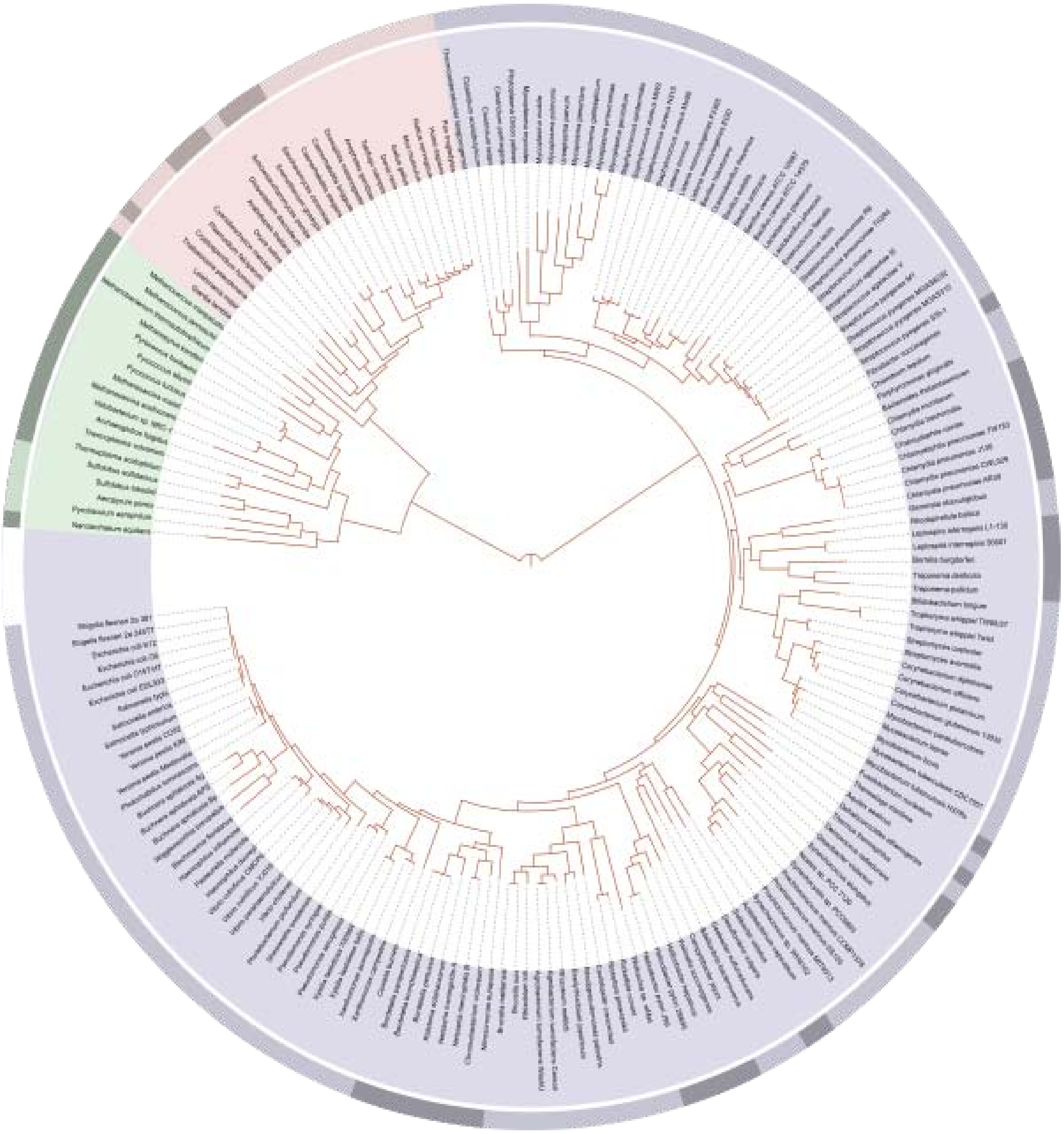
A cladogram tree has an implicit time axis, with time running forward from the base of the tree to the leaves of the tree. If the approximate date (for example, expressed as millions of years ago) of all the evolutionary forks were known, those dates could be captured in the cladogram. Thus, the time 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

Hominidae cannot be placed in any other monophyletic family. They also argue that paraphyletic taxa provide information about significant changes in organisms' morphology, ecology, or life history – in short, that both paraphyletic groups and clades are valuable notions with separate purposes.

Complexity of the Tree of Life

The cladistic tree of life is a fractal:

"The tree of life is inherently fractal-like in its complexity, ... Look closely at the 'lineage' of a phylogeny ... and it dissolves into many smaller lineages, and so on, down to a very fine scale."

The overall shape of a dichotomous (bifurcating) tree is recursive; as a viewpoint zooms into the tree of life, the same type of tree appears no matter what the scale. When extinct species are considered (both known and unknown), the complexity and depth of the tree can be very large. Moreover the tree continues to recreate itself by bifurcation, a series of events called fractal evolution. Every single speciation event, including all the species that are now extinct, represents an additional fork on the hypothetical, complete cladogram of the tree of life.

The tree of life is a quasi-self-similar fractal; that is, the deep reconstruction is not as regular as the shallow reconstruction. By shallow Mishler means the most recent branching toward and at the tips, and by deep the more ancient branches further back, which are harder to reconstruct and are missing unknown extinct lines. In the shallow part of the tree, branching events are relatively regular; it is often possible to estimate the times between them. In the deep part of the tree, "homology assessments" are "difficult" and the times vary widely. At this level Eldredge's and Gould's punctuated equilibrium applies, which hypothesizes long periods of stability followed by punctuations of rapid speciation, based on the fossil record.

PhyloCode approach to naming species

A formal code of phylogenetic nomenclature, the PhyloCode, is currently under development. It is intended for use by both those who would like to abandon Linnaean taxonomy and those who would like to use taxa and clades side by side. In several instances it has been employed to clarify uncertainties in Linnaean systematics so that in combination they yield a taxonomy that unambiguously places problematic groups in the evolutionary tree in a way that is consistent with current knowledge.

Example

For example, Linnaean taxonomy contains the taxon Tetrapoda, defined morphologically as vertebrates with four limbs (as well as animals with four-limbed ancestors, such as snakes), which is often given the rank of superclass, and divides into the classes Amphibia, Reptilia, Aves, Mammalia.

Phylogenetic nomenclature also contains the taxon Tetrapoda, whose living members can be classified phylogenically as "the clade defined by the common ancestor of amphibians and mammals", or more precisely the clade defined by the common ancestor of a specific amphibian and mammal (or bird or snake). This definition gives us the Crown group tetrapods (or Crown-Tetrapoda). A few primitive four legged ancestors (the Ichthyostegalia) fall outside Crown-Tetrapoda. An alternative is to define tetrapoda as all animals more closely related to mammals than to lungfish (our nearest living non-tetrapod relatives). In this definition, the ichthyostegalians are included, together with a host of fossil animals usually classed as crossopterygian fish. This wider definition is termed Pan-Tetrapoda. A third option is to define Tetrapoda according to their apomorphy (their unique trait, i.e. having legs rather than fins), a definition that yield the same group as the Linnaean taxon.

Non of the phylogenetic taxa as described above have a rank, and neither do its subtaxa. All the subclades are contained within one another. The clades are not divided into several non-overlapping taxa (as in traditional taxonomy), rather the clade is split into two clades at the first branching, a process repeated throughout. With regards to the traditional classes, Aves and Mammalia are subclades, contained in the subclade Amniota, while Reptilia and Amphibia are paraphyletic taxa, not clades. Instead of classifying non-mammalian, non-avian amniotes as reptiles, Amniota is divided into the two clades Sauropsida (which contains birds and all living amniotes other than mammals, including all living traditional reptiles) and Theropsida (mammals and the extinct mammal-like reptiles). Similarly, Amphibia can be split into the Batrachomorpha (fossil amphibians more closely related to modern amphibians) and Reptiliomorpha, the latter of which the amiotes is a sub-clade. Ichthyostegalians and other Stem-tetrapods represent sister groups from splits predating the Batrachomorpha/Reptilopmorpha split.

Summary of advantages of phylogenetic nomenclature

Proponents of phylogenetic nomenclature enumerate key distinctions between phylogenetic nomenclature and Linnaean taxonomy as follows:

Phylogenetic Nomenclature	**Linnaean Taxonomy**
Handles arbitrarily deep trees.	Often must invent new level names (such as superorder, suborder, infraorder, parvorder, magnorder) to accommodate new discoveries. Biased towards trees about 4 to 12 levels deep.
Discourages naming or use of groups that are not monophyletic	Acceptable to name and use paraphyletic groups
Primary goal is to reflect actual process of evolution	Primary goal is to group species based on morphological similarities

Assumes that the shape of the tree will change frequently with new discoveries	New discoveries often require renaming or releveling of Classes, Orders, and Kingdoms

Summary of criticisms of phylogenetic nomenclature

Critics of phylogenetic nomenclature include Ashlock, Mayr, and Williams. Some of their criticisms include:

Phylogenetic Nomenclature	**Linnaean Taxonomy**
Limited to entities related by evolution or ancestry	Supports groupings without reference to evolution or ancestry
Does not include a process for naming species	Includes a process for giving unique names to species
Clade definitions emphasize ancestry at the expense of descriptive characteristics	Taxa definitions based on tangible characteristics
Ignores sensible, clearly defined paraphyletic groups such as reptiles	Permits clearly defined groups such as reptiles
Difficult to determine if a given species is in a clade or not (e.g. if clade X is defined as "most recent common ancestor of A and B along with its descendants", then the only way to determine if species Y is in the clade is to perform a complex evolutionary analysis)	Straightforward process to determine if a given species is in a taxon or not
Limited to organisms that evolved by inherited traits; not applicable to organisms that evolved via complex gene sharing or lateral transfer	Applicable to all organisms, regardless of evolutionary mechanism

Application to other disciplines

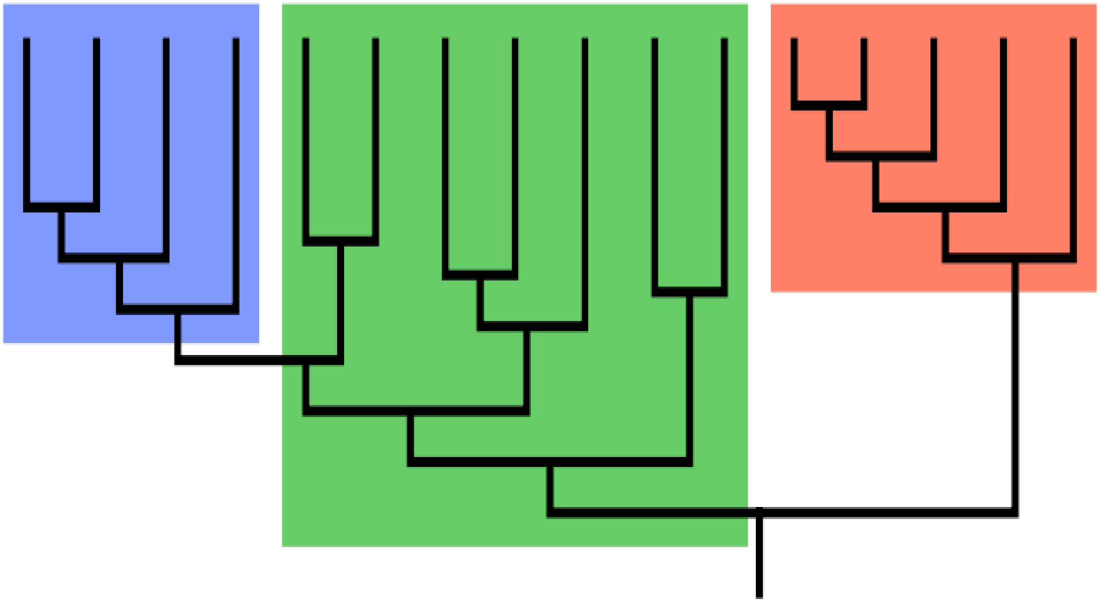
The comparisons used to acquire data on which cladograms can be based are not limited to the field of biology. Any group of individuals or classes, hypothesized to have a common ancestor, and to which a set of common characteristics may or may not apply, can be compared pairwise. Cladograms can be used to depict the hypothetical descent relationships within groups of items in many different academic realms. The only requirement is that the items have characteristics that can be identified and measured.

Recent attempts to use cladistic methods outside of biology address the reconstruction of lineages in:

- Anthropology and archeology. Compares cultures or artifacts using groups of cultural traits or artifact features.
- Linguistics. Compares languages using groups of linguistic features.
- Textual criticism or Stemmatics. Compares manuscripts of the same work (original lost) using groups of distinctive copying errors.
- Ethology. Compares animal species using behavioral traits presumed hereditary.

Chapter- 2

Clade

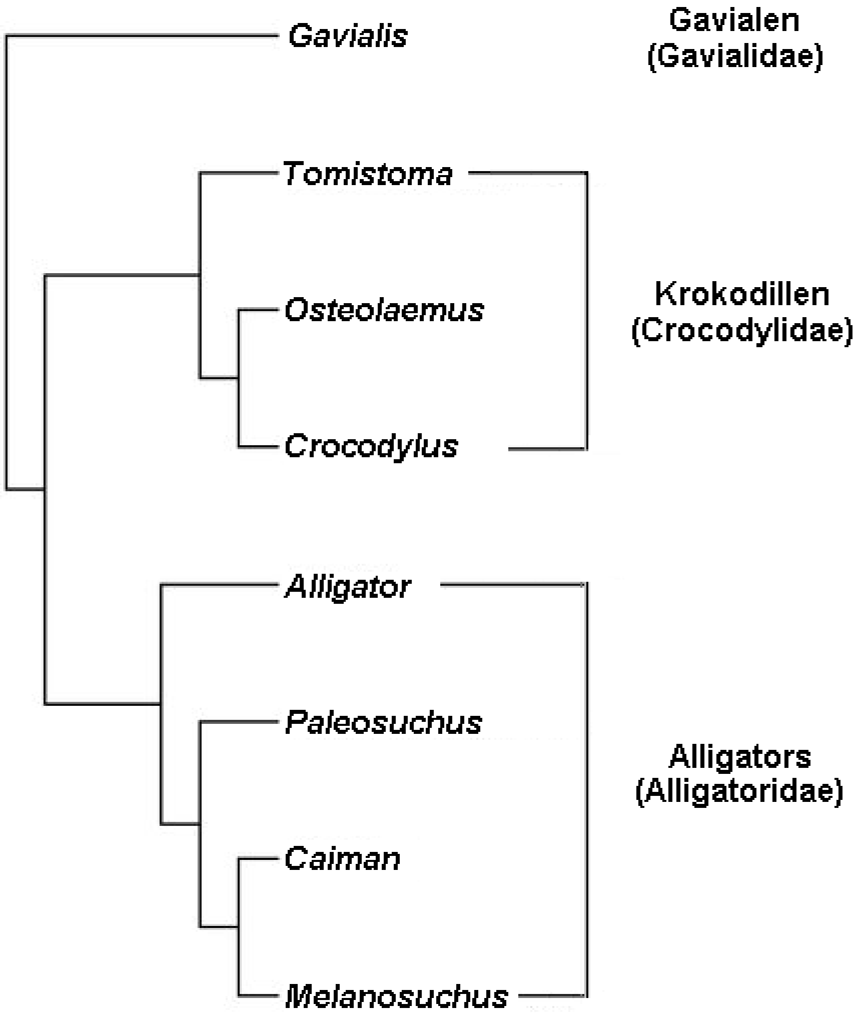


Cladogram (family tree) of a biological group. The red and blue boxes represent *clades* (i.e., complete branches). The green box is not a clade, but rather represents an *evolutionary grade*, an incomplete group, because the blue clade descends from it, but is excluded.

A **clade** is a group consisting of an organism and all its descendants. In the terms of biological systematics, a clade is a single "branch" on the "tree of life". The idea that such a "natural group" of organisms should be grouped together and given a taxonomic name is central to biological classification. In cladistics (which takes its name from the term), clades are the only acceptable units.

The term was coined in 1958 by English biologist Julian Huxley.

Definitions



A cladogram of crocodiles, a visual representation of their relationship

Clade and ancestor

A clade is termed monophyletic, meaning it contains one ancestor which can be an organism, population, or species and all its descendants. The term clade refers to the grouping of the ancestor and its living and/or deceased descendants together. The ancestor can be a theoretical or actual species.

Clade definition

Three methods of defining clades are featured in phylogenetic nomenclature: node-, stem-, and apomorphy-based:

- In node-based definition, clade name A refers to the *least inclusive* clade containing taxa (or specimens) X, Y, etc., and their common ancestor. The ancestor is the branch point, or *node*.
- In stem-based definition, A refers to the *most inclusive* clade containing X, Y, etc., and their common ancestor, down to where Z branches off below A. Taxa are included between the node of A and down to (but not including) the branching point to Z; that is, the *stem* of A.
- In apomorphy-based definition, A refers to the clade identified by an apomorphy (a trait) found in X, Y, etc., and their common ancestor.

In Linnaean taxonomy, clades are defined by a set of traits (apomorphies) unique to the group. This system is basically similar to the apomorphy-based clades of phylogenetic nomenclature. The difference is one of weight: While phylogenetic nomenclature bases the group on an ancestor with a certain trait, Linnaean taxonomy uses the traits themselves to define the group.

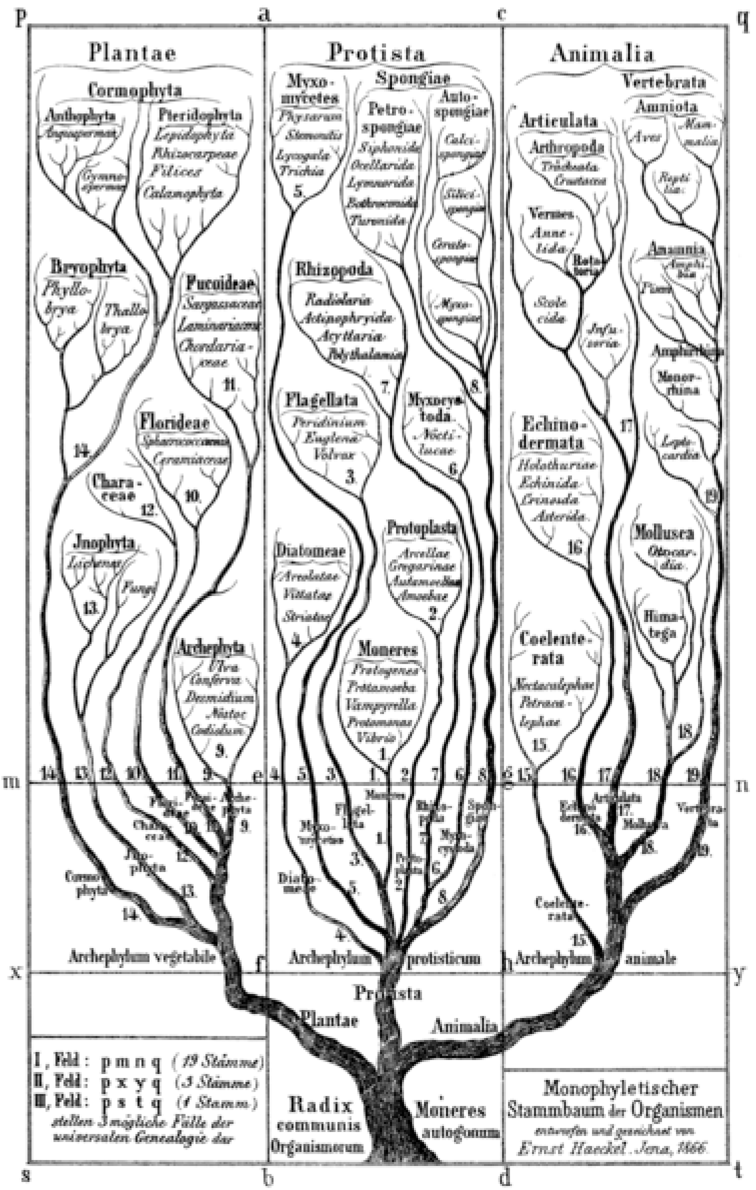
Clades as constructs

In cladistics, the clade is a hypothetical construct based on experimental data. Clades are found using multiple (sometimes hundreds) of traits from a number of species (or specimens) and analysing them statistically to find the most likely phylogenetic tree for the group. Although similar in some ways to a biological classification of species, the method is statistical and more open to scrutiny than traditional methods. Although taxonomists use clades as a tool in classification where feasible, the taxonomic "tree of life" is not the same as the cladistic. The traditional genus, family, etc. names are not necessarily clades; though they will often be.

Clade names

In Linnaean systematics, the various groups are ordered into a series of taxonomic ranks (the familiar order, family etc.). These ranks will by convention dictate the ending to names for some groups. Clades do not by their nature fit this scheme, and no such restriction exists as to their names in cladistics. There is however a convention for naming more or less inclusive groups, which are given prefixes like *crown-* or *pan.*

Taxonomy and systematics



Early phylogenetic tree by Haeckel, 1866

The idea of a "clade" did not exist in pre-Darwinian Linnaean taxonomy, which was based only on morphological similarities between organisms – although it happens that many of the better known animal groups in Linnaeus' original Systema Naturae (notably among the vertebrate groups) do represent clades. With the publication of Darwin's theory of evolution in 1859, taxonomy gained a theoretical basis, and the idea was born that groups used in a system of classification should represent branches on the evolutionary tree of life. In the century and a half since then, taxonomists have worked to make the taxonomic system reflect evolution. However, partly because the Tree of Life branches rather unevenly, the hierarchy of the Linnaean system does not always lend itself well to representing clades. The result is that when it comes to naming, cladistics and Linnaean taxonomy are not always compatible. In particular, higher level taxa in

Linnaean taxonomy often represent evolutionary grades rather than clades, resulting in groups made up of clades where one or two sub-branches have been excluded. Typical examples include bony fishes, which include the ancestor of tetrapods, and reptiles, ancestral to both birds and mammals.

In phylogenetic nomenclature, clades can be nested at any level, and do not have to be slotted into a small number of ranks in an overall hierarchy. In contrast, the Linnaean units of "order", "class" etc. must be used when naming a new taxon. As there are only seven formal levels to the Linnaean system (species being the lowest), only a finite number of sub- and super-units can be created. In order to be able to use the full complexity of taxonomic trees (cladograms) in an area with which they are very familiar, some researchers have opted to dispense with ranks all together, instead using clade names without Linnaean ranks. The reason for preferring one system over the other is partly one of application: cladistic trees give details, suitable for specialists; the Linnaean system gives a well ordered overview, at the expense of details of the phylogenetic tree.

In a few instances, the Linnaean system has actually impeded our understanding of the phylogeny and broad evolutionary patterns. The best known example is the interpretation of the strange fossils of the Burgess Shale and the subsequent idea of a "Cambrian Explosion" With the application of cladistics, and the rejection of any significance of the concept of phyla, the confusion of the late 20th century over the Burgess animals has been resolved. It appears there never was an "explosion" of major bauplans with subsequent extinctions. The seemingly weird critters themselves have been found to be representatives of a group, the Lobopodia, that includes arthropods, water bears and velvet worms.

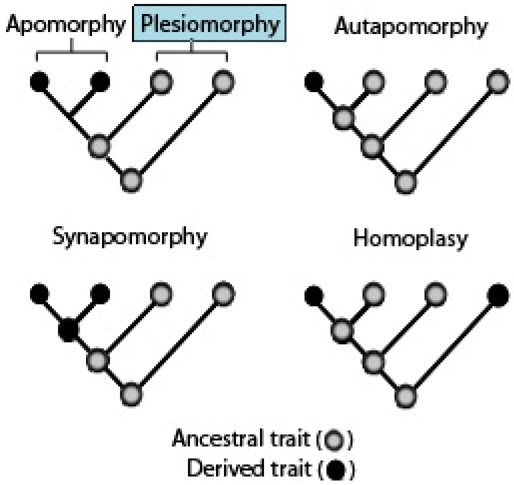
In most instances the two systems are not at odds, however. The cladistic statement, that the clade Lobopodia contains (among others) the Arthropoda, Tardigrada and Onychophora, is factually identical to the Linnaean evolutionary statement that the group Lobopodia is ancestral to the phyla Arthropoda, Tardigrada and Onychophora. The difference is one of semantics rather than phylogeny.

Chapter- 3

Terminology for Characters

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

Symplesiomorphy



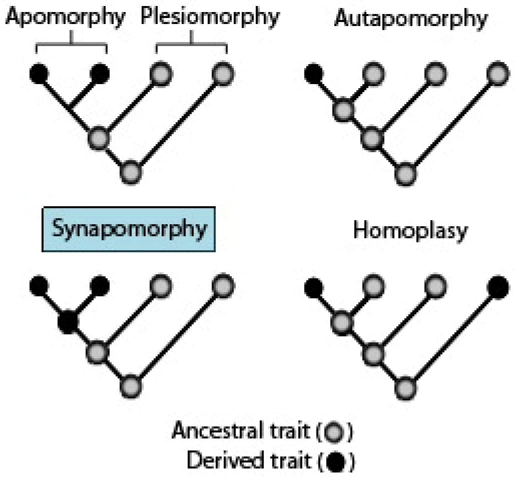
A cladogram showing the terminology used to describe different patterns of ancestral and derived character states.

In cladistics, a **symplesiomorphy** or **symplesiomorphic character** is a trait which is shared (a symmorphy) between two or more taxa, but which is also shared with other taxa which have an earlier last common ancestor with the taxa under consideration. They are therefore not an indication that the taxa be considered more closely related to each other than to the more distant taxa, as all share the more primitive trait. A close phylogenetic relationship, that the taxa form a certain clade to the exclusion of certain other taxa, can only be shown by the discovery of synapomorphies: shared traits that have originated with the last common ancestor of the taxa considered, or at least in the branch, not including the taxa to be excluded, leading to it.

The concept of the symplesiomorphy shows the danger of grouping species together on the basis of general morphologic or genetic similarity, without distinguishing between resemblances caused by either primitive or derived traits. This phenetic method of analysis was common before cladistics became popular in the 1980s.

A famous example is pharyngeal gill breathing in bony and cartilaginous fishes. The former are more closely related to Tetrapoda (terrestrial vertebrates, which evolved out of a clade of bony fishes) that breathe via their skin or lungs, rather than to the sharks, rays, *et al.*. Their kind of gill respiration is shared by the "fishes" because it was present in their common ancestor and lost in the other living vertebrates.

Synapomorphy



A cladogram showing the terminology used to describe different patterns of ancestral and derived character states.

In cladistics, a **synapomorphy** or **synapomorphic character** is a trait that is shared ("symmorphy") by two or more taxa and their most recent common ancestor, whose ancestor in turn does not possess the trait. A synapomorphy is thus an apomorphy visible in multiple taxa, where the trait in question originates in their last common ancestor. The word "synapomorphy" is derived from the Greek words *σύν*, *syn* = with, in company with, together with; *ἀπό*, *apo* = away from; and *μορφή*, *morphe* = shape.

True synapomorphies usually uniquely characterise a given set of terminal groups, but this is not essential to the concept. Thus, if some descendants of a last common ancestor possess a synapomorphic trait, it is not strictly necessary that all of its descendants must possess the same trait.

Comparisons with other shared traits

A synapomorphy should not be confused with other types of shared traits:

- A synapomorphy is a shared trait found among two or more taxa and their most recent common ancestor, whose ancestor in turn does not possess the trait. An example is the halteres, the uniquely modified hind wings found in all families of winged Diptera. No other group of insects possesses similar structures. However, the fact that the trait is found exclusively in Diptera, to the exclusion of all other groups, is not essential in identifying the trait as a synapomorphy; rather, this fact makes its determination easier.

- A symplesiomorphy is a shared trait found among two or more taxa, but which is also found in taxa with an earlier common ancestor. An example of this is the five toes seen on the hind legs of rats and apes. This character-state originated very early in Tetrapoda and occurs in other tetrapod groups, e.g. in lizards. There is thus no indication that the group formed of rats and apes is a clade to the exclusion of these other groups.

- A homoplasy is a shared trait found among different taxa but not in their common ancestor (i.e., the same trait emerged in different taxa independently of each other). An example of this is homeothermy in birds and mammals. This trait is a derived character-state (in relation to poikilothermy, the character-state of the last common ancestor of both groups), which evolved independently in these two groups (or at least in the larger clades to which these groups belong).

Cladistic analyses

Synapomorphies are used to establish phylogenies in cladistic analyses. As such they are empirical data which can support a certain hypothesis that terminal groups form a clade (monophyletic group) together to the exclusion of certain other groups – whereas character-states that are shared, but also shared by other terminal groups descending from an earlier common ancestor, cannot be used to exclude these other groups. The latter character-states can consist of symplesiomorphies ("primitive" character-states having originated in the earlier common ancestor) or homoplasies (superficially similar but independently evolved derived character-states).

The key problem is to identify the polarity of the transformation series to which several character-states belong, i.e. to tell which character-state is apomorphic and which is plesiomorphic. Various criteria were used to polarise the transformation series in earlier cladistics; however in the recent two decades pattern criteria based on outgroup comparison have dominated the field.

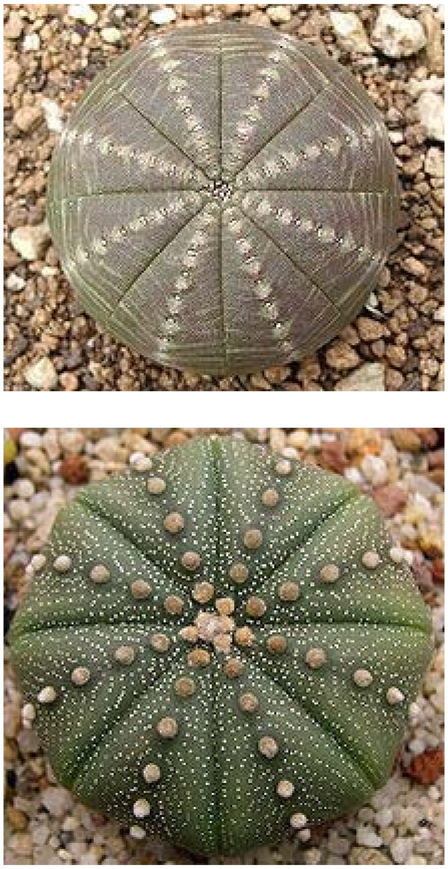
The concepts of apomorphy and plesiomorphy are relative to a certain level of generality. What counts as an apomorphy at one level of generality may well be a plesiomorphy at the other. For example, for rats and apes, the presence of mammary glands is a symplesiomorphy, but it is a synapomorphy for mammals in relation to tetrapods more broadly.

It is not essential to a synapomorphy that all members of a clade possess it; even if some would have secondarily lost the trait it could still be a synapomorphy of the clade as a whole. A character state that is a synapomorphy for a clade, but for lineages in this clade is a plesiomorphy that is altered in some lineages, is called **underlying synapomorphy**. If no crown group taxa are known, it is sometimes difficult to decide which character state is the underlying synapomorphy and which the autapomorphy that overlies it.

Clades are not defined by synapomorphies as such, though it is possible to define them by apomorphies in general.

Convergent evolution

Convergent evolution



These two succulent plant genera, *Euphorbia*

and *Astrophytum*, are only distantly
related, but have independently converged
on a very similar body form.

Convergent evolution describes the acquisition of the same biological trait in unrelated lineages.

The wing is a classic example of convergent evolution in action. Although their last common ancestor did not have wings, birds and bats do, and are capable of powered flight. The wings are similar in construction, due to the physical constraints imposed upon wing shape. Similarity can also be explained by shared ancestry, as evolution can only work with what is already there—thus wings were modified from limbs, as evidenced by their bone structure.

Traits arising through convergent evolution are termed analogous structures, in contrast to homologous structures, which have a common origin. Bat and pterodactyl wings are an example of analogous structures, while the bat wing is homologous to human and other mammal forearms, sharing an ancestral state despite serving different functions. Similarity in species of different ancestry that is the result of convergent evolution is called **homoplasy**. The opposite of convergent evolution is divergent evolution, whereby related species evolve different traits. On a molecular level, this can happen due to random mutation unrelated to adaptive changes. Convergent evolution is similar to, but distinguishable from, the phenomena of evolutionary relay and parallel evolution. Evolutionary relay describes how independent species acquire similar characteristics through their evolution in similar ecosystems at different times—for example the dorsal fins of extinct ichthyosaurs and sharks. Parallel evolution occurs when two independent species evolve together *at the same time in the same ecospace* and acquire similar characteristics—for instance extinct browsing-horses and paleotheres.

Causes

Similarity can also result if organisms occupy similar ecological niches—that is, a distinctive way of life. A classic comparison is between the marsupial fauna of Australia and the placental mammals of the Old World. The two lineages are clades—that is, they each share a common ancestor that belongs to their own group, and are more closely related to one another than to any other clade—but very similar forms evolved in each isolated population. Many body plans, for instance sabre-toothed cats and flying squirrels, evolved independently in both populations.

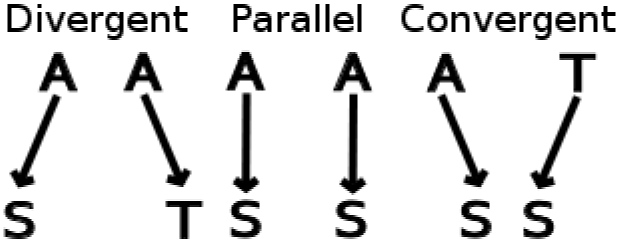
Distinction from re-evolution

In some cases, it is difficult to tell whether a trait has been lost then re-evolved convergently, or whether a gene has simply been 'switched off' and then re-enabled later. From a mathematical standpoint, an unused gene has a reasonable probability of remaining in the genome in a functional state for around 6 million years, but after 10 million years it is almost certain that the gene will no longer function.

Examples

One of the most famous examples of convergent evolution is the camera eye of cephalopods (e.g., squid), vertebrates (e.g., mammals) and cnidaria (e.g., box jellies). Their last common ancestor had at most a very simple photoreceptive spot, but a range of processes led to the progressive refinement of this structure to the advanced camera eye - with one subtle difference: The cephalopod eye is "wired" in the opposite direction, with blood and nerve vessels entering from the back of the retina, rather than the front as in vertebrates. The similarity of the structures in other respects, despite the complex nature of the organ, illustrates how there are some biological challenges (vision) that have an optimal solution.

Parallel vs. convergent evolution



Evolution at an amino acid position. In each case, the left-hand species changes from incorporating alanine (A) at a specific position within a protein in a hypothetical common ancestor deduced from comparison of sequences of several species, and now incorporates serine (S) in its present-day form. The right-hand species may undergo divergent, parallel, or convergent evolution at this amino acid position relative to that of the first species.

For a particular trait, proceeding in each of two lineages from a specified ancestor to a later descendant, parallel and convergent evolutionary trends can be strictly defined and clearly distinguished from one another. When both descendants are similar in a particular respect, evolution is defined as parallel if the ancestors considered were also similar, and convergent if they were not.

When the ancestral forms are unspecified or unknown, or the range of traits considered is not clearly specified, the distinction between parallel and convergent evolution becomes more subjective. For instance, the striking example of similar placental and marsupial forms is described by Richard Dawkins in *The Blind Watchmaker* as a case of convergent evolution, because mammals on each continent had a long evolutionary history prior to the extinction of the dinosaurs under which to accumulate relevant differences. Stephen Jay Gould describes many of the same examples as parallel evolution starting from the common ancestor of all marsupials and placentals. Many evolved similarities can be described in concept as parallel evolution from a remote ancestor, with the exception of those where quite different structures are co-opted to a similar function. For example, consider *Mixotricha paradoxa*, a microbe that has assembled a system of rows of apparent cilia and basal bodies closely resembling that of ciliates but that are actually smaller symbiont micro-organisms, or the differently oriented tails of fish and whales. On the converse, any case in which lineages do not evolve together at the same time in the same ecospace might be described as convergent evolution at some point in time.

The definition of a trait is crucial in deciding whether a change is seen as divergent, or as parallel or convergent. In the image above, note that, since serine and threonine possess similar structures with an alcohol side-chain, the example marked *"divergent"* would be termed *"parallel"* if the amino acids were grouped by similarity instead of being considered individually. As another example, if genes in two species independently become restricted to the same region of the animals through regulation by a certain transcription factor, this may be described as a case of parallel evolution - but examination of the actual DNA sequence will probably show only divergent changes in individual base-pair positions, since a new transcription factor binding site can be added in a wide range of places within the gene with similar effect.

A similar situation occurs considering the homology of morphological structures. For example, many insects possess two pairs of flying wings. In beetles, the first pair of wings is hardened into wing covers with little role in flight, while in flies the second pair of wings is condensed into small halteres used for balance. If the two pairs of wings are considered as interchangeable, homologous structures, this may be described as a parallel reduction in the number of wings, but otherwise the two changes are each divergent changes in one pair of wings.

Similar to convergent evolution, evolutionary relay describes how independent species acquire similar characteristics through their evolution in similar ecosystems, but not at the same time (dorsal fins of sharks and ichthyosaurs).

Significance

The degree to which convergence affects the products of evolution is the subject of a popular controversy. In his book *Wonderful Life*, Stephen Jay Gould argues that if the tape of life were re-wound and played back, life would have taken a very different course. Simon Conway Morris counters this argument, arguing that convergence is a dominant force in evolution, and that, since the same environmental and physical constraints act on all life, there is an "optimum" body plan that life will inevitably evolve toward, with evolution bound to stumble upon intelligence - a trait of primates, crows, and dolphins - at some point. Convergence is difficult to quantify, so progress on this issue may require exploitation of engineering specifications (e.g., of wing aerodynamics) and comparably rigorous measures of "very different course" in terms of phylogenetic (molecular) distances.

Cultural convergence

The term *convergence* is also used to describe phenomena in the theory of cultural evolution.

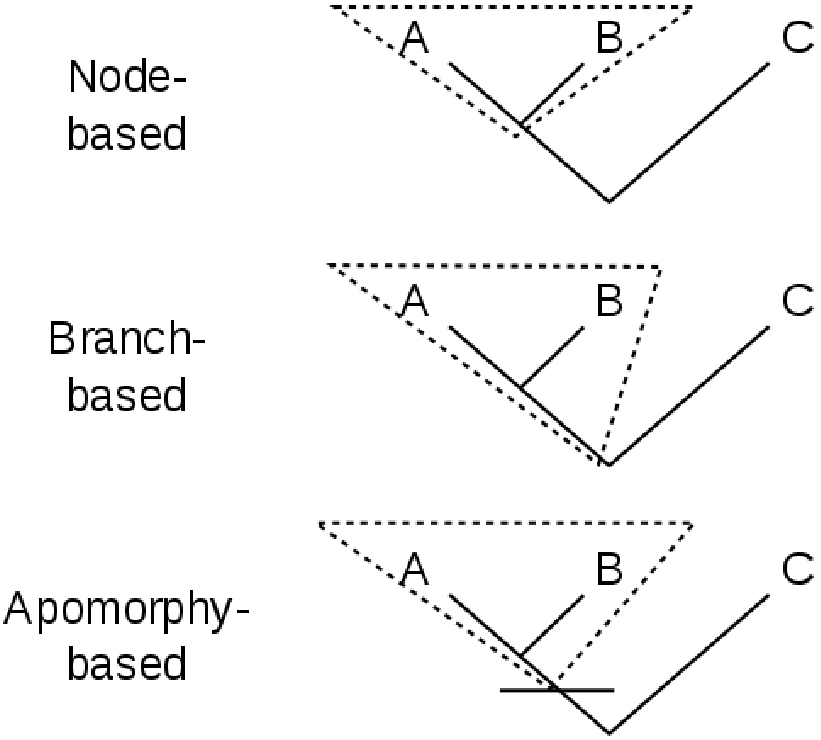
Chapter- 4

Phylogenetic Nomenclature

Phylogenetic nomenclature (PN) or **phylogenetic taxonomy** is an alternative to rank-based nomenclature, applying definitions from cladistics (or *phylogenetic systematics*). Its two defining features are the use of *phylogenetic definitions* of biological taxon names, and the *lack of obligatory ranks*. It is currently not regulated, but the *PhyloCode* (*International Code of Phylogenetic Nomenclature*) is intended to regulate it once implemented.

The terms **cladism** and **cladist** were first introduced by Ernst W. Mayr in 1965. They sometimes refer to cladistics as a whole.

Definitions



Types of phylogenetic definitions. The horizontal bar indicates the evolution of the apomorphy mentioned in the apomorphy-based definition. "Branch-based" was formerly called "stem-based"; the term was changed to avoid confusion with the term "stem group" which means total clade minus crown clade.

Under the rank-based codes of biological nomenclature, names themselves do not have definitions, but are instead usually linked to a type. Some biologists have claimed that this is unsatisfactory and that instability in nomenclature should only reflect instability of our knowledge of phylogeny, not instability in subjective opinions about which ranks should be given to which groups. Phylogenetic nomenclature, on the other hand, uses phylogenetic definitions to tie a name to a clade in such a manner that the meaning of the name is objective under any phylogenetic hypothesis, thus preventing splitting and lumping (unless definitions are changed in the process, which will be allowed under the *PhyloCode* only under carefully restricted circumstances).

Traditionally, groups named in phylogenetic nomenclature are usually monophyletic-that is, they define a natural group made up of all descendants of a single common ancestor. However, it is also possible to create phylogenetic definitions for the names of paraphyletic taxa . Assuming Mammalia and Aves are defined, Reptilia could be defined as "the most recent common ancestor of birds and mammals and all its descendants except birds and mammals". This includes taxa that are not currently named and even taxa that cannot be named under the rank-based codes without seriously disrupting existing classifications, such as "all organisms that share a more recent common ancestor with *Homo sapiens* than with birds and plesiomorphically keep laying eggs". Names of polyphyletic taxa could be defined by referring to the sum of two or more clades or paraphyletic taxa .

Philosophy

Rank-based and phylogenetic nomenclature differ in philosophical outlook. Rank-based nomenclature is linked to classification: it starts with the known species (or subspecies or varieties or even individuals), waits for an act of classification to group them into larger taxa, and then asks how to name these taxa. Phylogenetic nomenclature, on the other hand, starts with the phylogenetic tree of life (as hypothesized by the science of phylogenetics) and asks how and where to tie labels to its branches. Furthermore, phylogenetic nomenclature follows the nomenclature in other sciences in trying to define its terms as precisely as possible, while rank-based nomenclature deliberately keeps its definitions incomplete; for example, Principle 2 of the ICZN is "Nomenclature does not determine the inclusiveness or exclusiveness of any taxon". This is the result of a third philosophical difference: users of rank-based nomenclature commonly start from a name and ask for its meaning (in other words: which taxon the name should be applied to), while users of phylogenetic nomenclature tend to start from a clade and ask what to call it.

Lack of ranks

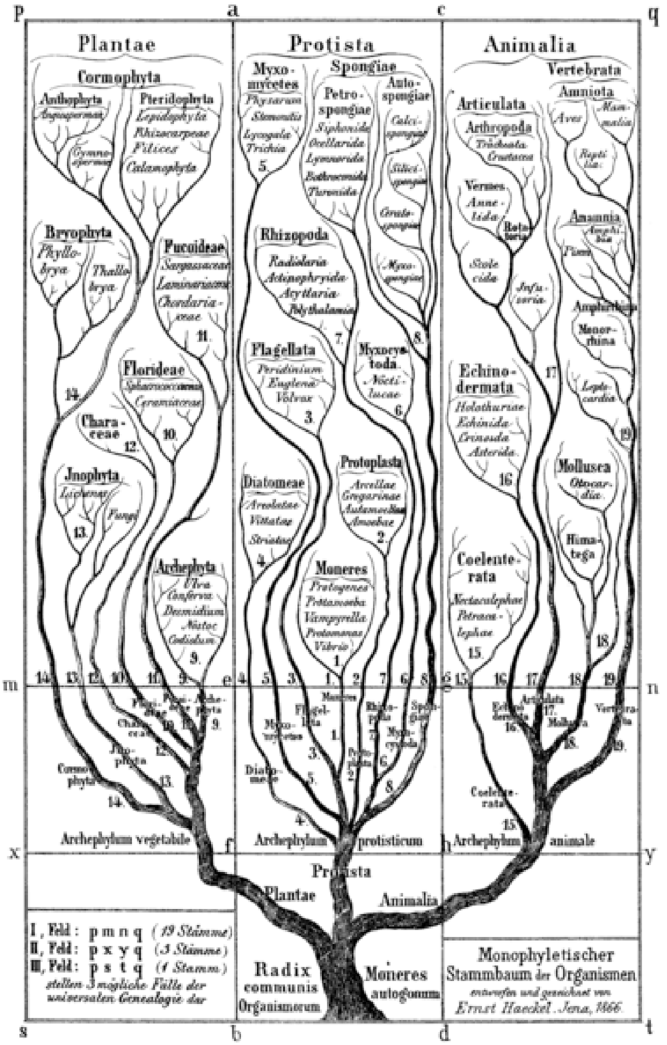
The current codes of biological nomenclature stipulate that taxa cannot be given a valid name without being given a rank. However, the number of generally recognized ranks is limited. Gauthier *et al.* (1988) showed that a classification which uses the common array of ranks, but includes Aves within Reptilia and keeps Reptilia at its traditional rank of class, is forced to demote Aves to the rank of genus, despite the ~ 12,000 known species of extant and extinct birds that would have to be incorporated into this one genus. To reduce this problem, Patterson and Rosen (1977) suggested nine new ranks between family and superfamily in order to be able to classify a clade of herrings, and McKenna and Bell (1997) introduced a large array of new ranks in order to cope with the diversity of Mammalia.

The current codes also each have rules saying that names must have certain endings if they are applied to taxa that have certain ranks. When a taxon changes rank from one classification to another, its name must change its suffix. Ereshefsky (1997:512) stated:

The Linnaean rule of assigning rank-specific suffices [sic] gives rise to even more confusing cases. Simpson (1963, 29–30) and Wiley (1981, 238) agree that the genus *Homo* belongs to a particular taxon. They disagree, however, on that taxon's rank. Acting in accord with the Linnaean system, they attach different suffixes to the root *Homini* [actually Homin-] and give the taxon in question different names: Wiley calls it 'Hominini' [tribe rank] and Simpson calls it 'Hominidae' [family rank]. Their disagreement does not stop there. Wiley believes that the taxon just cited is a part of a more inclusive taxon which is a family. Using the root *Homini*, and following the rules of the Linnaean system [more precisely, the zoological code], he names the more inclusive taxon 'Hominidae.' So for Wiley and Simpson, the name 'Hominidae' refers to two different taxa. In brief, the Linnaean system causes Wiley and Simpson to assign different names to what they agree is the same taxon, and it causes them to give the same name to what they agree are different taxa.

In phylogenetic nomenclature, ranks have no bearing on the spelling of taxon names. Ranks are, however, not altogether forbidden in phylogenetic nomenclature. They are merely decoupled from nomenclature: they do not influence which names can be used, which taxa are associated with which names, and which names can refer to nested taxa (e.g.).

History



"Monophyletic tree of organisms". Ernst Haeckel: *Generelle Morphologie der Organismen,* etc. Berlin, 1866.

Ultimately, phylogenetic nomenclature is a result of Darwin's discovery that the diversity and history of life is best represented in tree-shaped diagrams. This discovery immediately led to changes in the existing classifications. For example, John Hogg proposed the term Protoctista in 1860 for organisms that did not seem closely related to either animals or plants. In 1866, the controversial biologist Ernst Haeckel for the first time reconstructed a single tree of all life (see figure) and immediately proceeded to translate it into a classification. This classification was rank-based, in accordance with the only code of biological nomenclature that existed at the time, but did not contain taxa that Haeckel considered polyphyletic; in it, Haeckel introduced the rank of phylum which carries a connotation of monophyly in its name.

Ever since it has been debated in which ways and to what extent the phylogeny of life should be used as a basis for its classification, with views ranging from "numerical taxonomy" (phenetics) over "evolutionary taxonomy" (gradistics) to "phylogenetic systematics" (cladistics – today, the term "cladistics" is only used for the method of phylogeny reconstruction, but its inventor, Willi Hennig, regarded this method as a mere tool for the purpose of classification). From the 1960s onwards, rankless classifications were occasionally proposed, but in general the principles of rank-based nomenclature were used by all three schools of thought.

Most of the basic tenets of phylogenetic nomenclature (lack of obligatory ranks, and something close to phylogenetic definitions) can, however, be traced to 1916, when Edwin Goodrich interpreted the name Sauropsida, erected 40 years earlier by Huxley, to include the birds (Aves) as well as *part of* Reptilia, and coined the new name Theropsida to include the mammals as well as another *part of* Reptilia, but did not give them ranks, and treated them exactly as if they had what would today be termed branch-based definitions, using neither contents nor diagnostic characters to decide whether a given animal should belong to Theropsida, Sauropsida, or something else once its phylogenetic position was agreed upon. Goodrich also opined that the name Reptilia should be abandoned once the phylogeny of the reptiles would be better known. The lack of compatibility of his scheme with the existing rank-based classifications (despite agreement on the phylogeny in all but details), and the lack of a method of phylogenetics at this time, are the most likely reasons why Goodrich's suggestions were largely ignored.

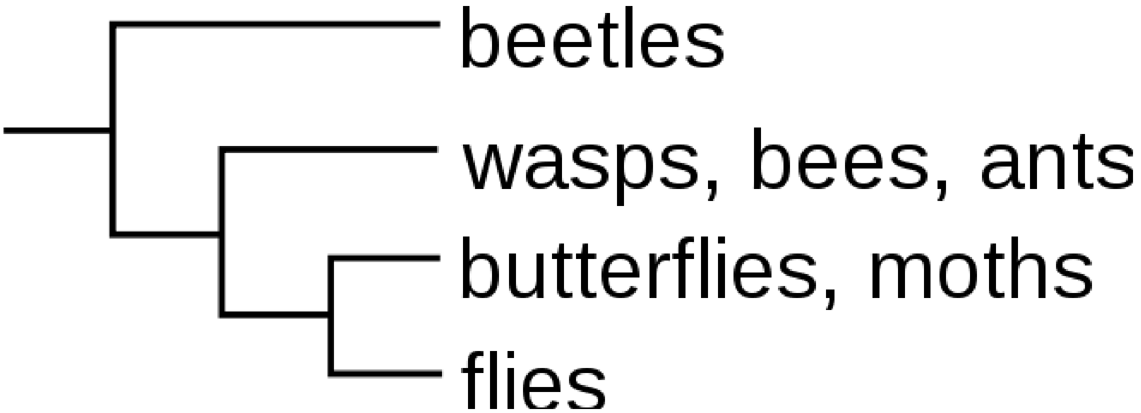
The principle that only clades (monophyletic taxa – an ancestor plus all its descendants) should be formally named became popular in the second half of the 20th century. It spread together with the methods for discovering clades (cladistics) and is an integral part of phylogenetic systematics (see above). At the same time, it became apparent that the obligatory ranks that are part of the traditional systems of nomenclature produced problems. Some authors suggested abandoning them altogether, starting with Willi Hennig's abandonment of his earlier proposal to define ranks as geological age classes.

The origin of phylogenetic nomenclature can be dated to 1986, when Jacques Gauthier used phylogenetic definitions for the first time in a published work. Theoretical papers outlining the principles of phylogenetic nomenclature, as well as further publications containing applications of phylogenetic nomenclature (mostly to vertebrates), soon followed.

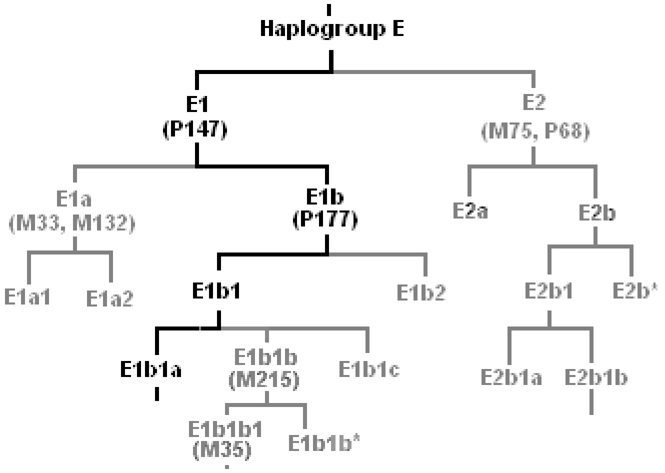
In an attempt to avoid a schism in the biologist community, "Gauthier suggested to two members of the ICZN to apply formal taxonomic names ruled by the zoological code only to clades (at least for supraspecific taxa) and to abandon Linnean ranks, but these two members promptly rejected these ideas" (Laurin, 2008: 224). This led him, Kevin de Queiroz, and the botanist Philip Cantino to start drafting their own code of nomenclature, the *PhyloCode*, for regulating phylogenetic nomenclature.

Chapter- 5

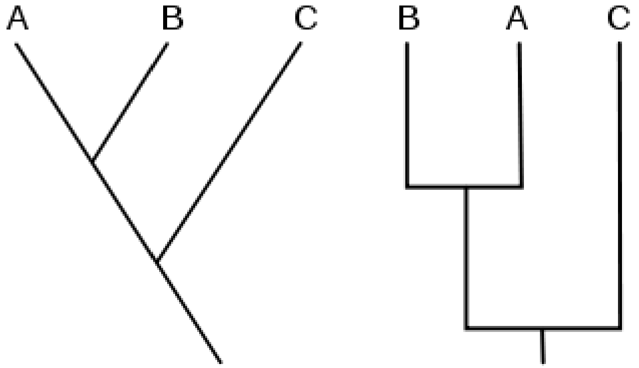
Cladogram



A horizontal cladogram, with the ancestor (not named) to the left



A vertical cladogram, with the ancestor at the top



Two vertical cladograms, the ancestor at the bottom

A **cladogram** is a diagram used in cladistics which shows ancestral relations between organisms, to represent the evolutionary tree of life. Although traditionally such cladograms were generated largely on the basis of morphological characters, DNA and RNA sequencing data and computational phylogenetics are now very commonly used in the generation of cladograms.

Generating a cladogram

A greatly simplified procedure for generating a cladogram is:

1. Gather and organize data
2. Consider possible cladograms
3. Select best cladogram

Step 1: gather and organize data

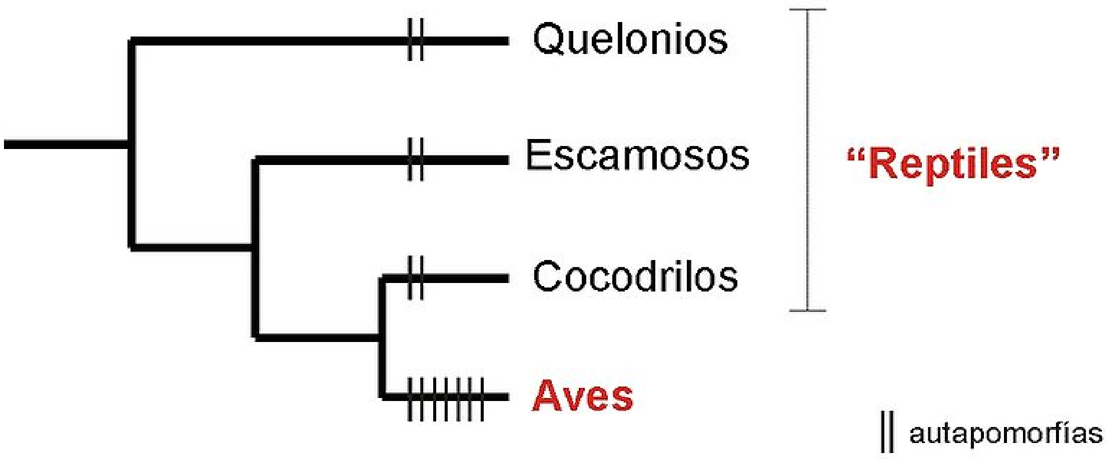
A cladistic analysis begins with the following data:

- a list of taxa (for example, species) to be organized
- a list of characteristics to be compared
- for each taxon, the value of each of the listed characteristics or *character states*

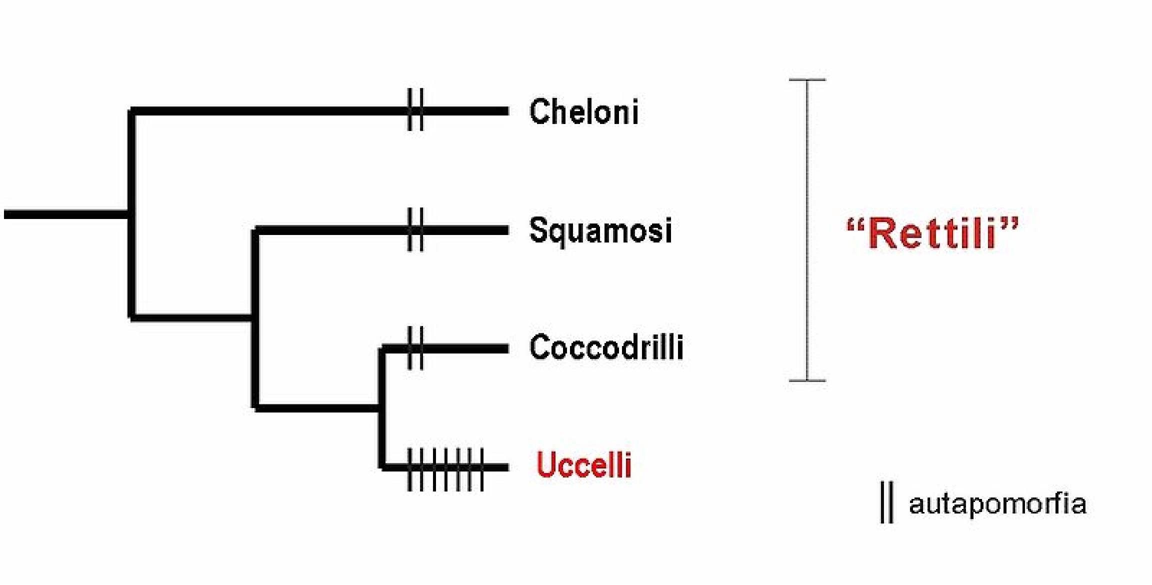
For example, if analyzing 20 species of birds, the data might be:

- the list of the 20 species
- characteristics such as genome sequence, skeletal anatomy, biochemical processes, and feather coloration
- for each of the 20 species, its particular genome sequence, skeletal anatomy, biochemical processes, and feather coloration

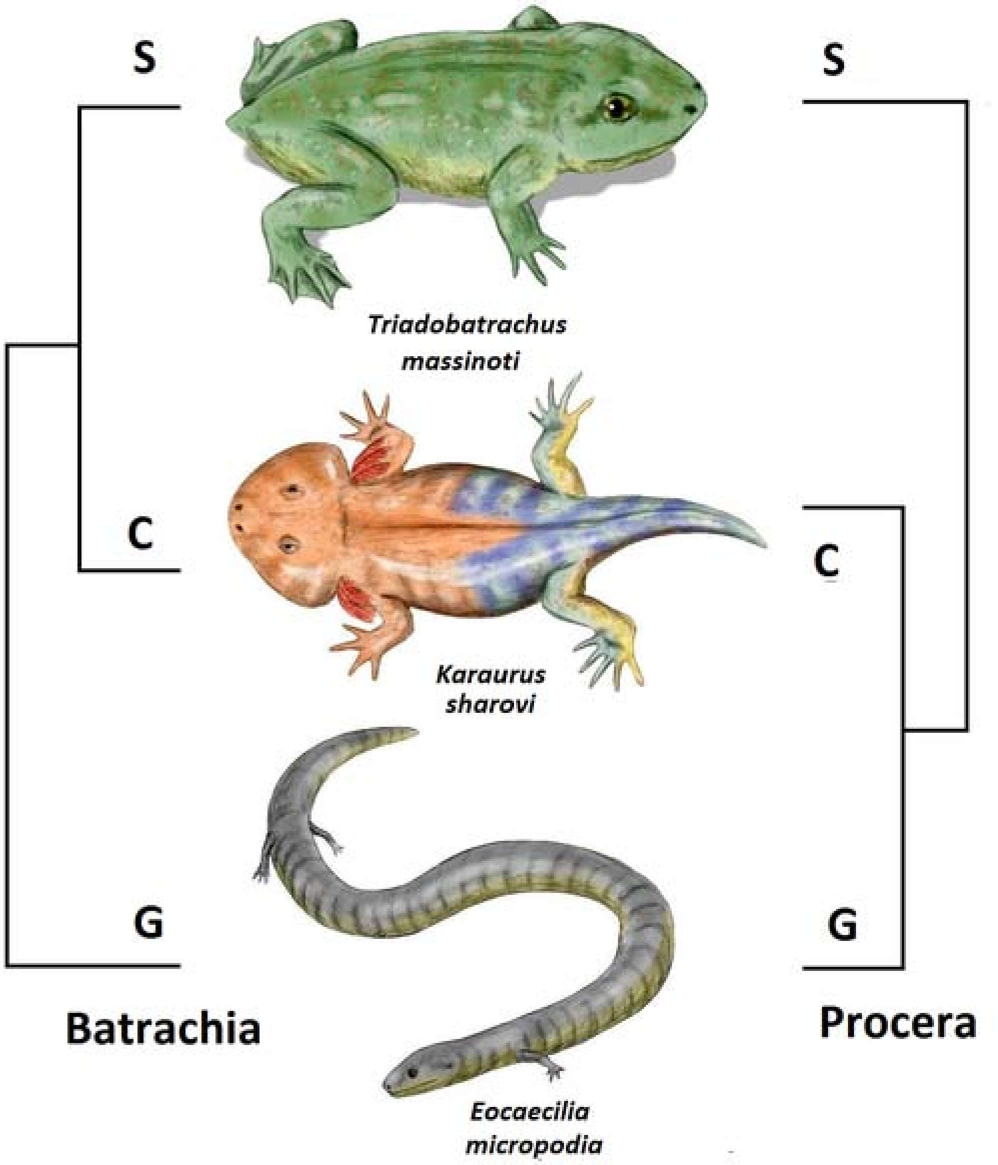
All the data are then organized into a "taxon-character matrix", which is the base to perform phylogenetic analysis.



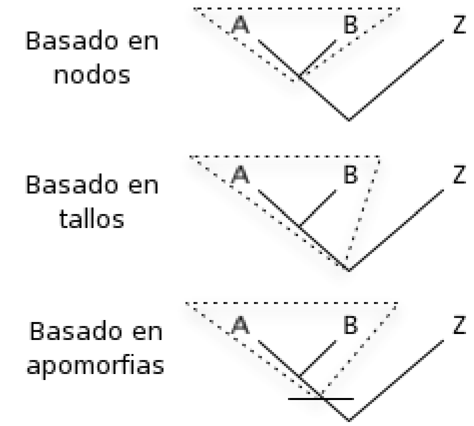
Aminiotas



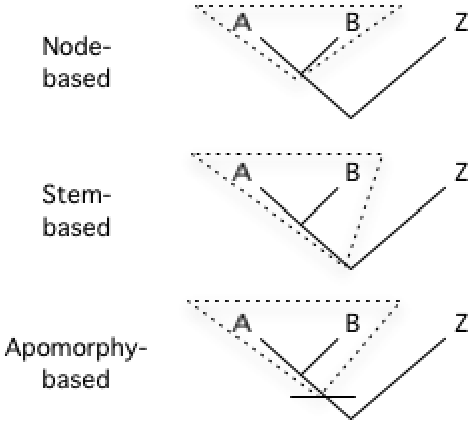
Amnioti



Batrachia-Procera



Clade types-es

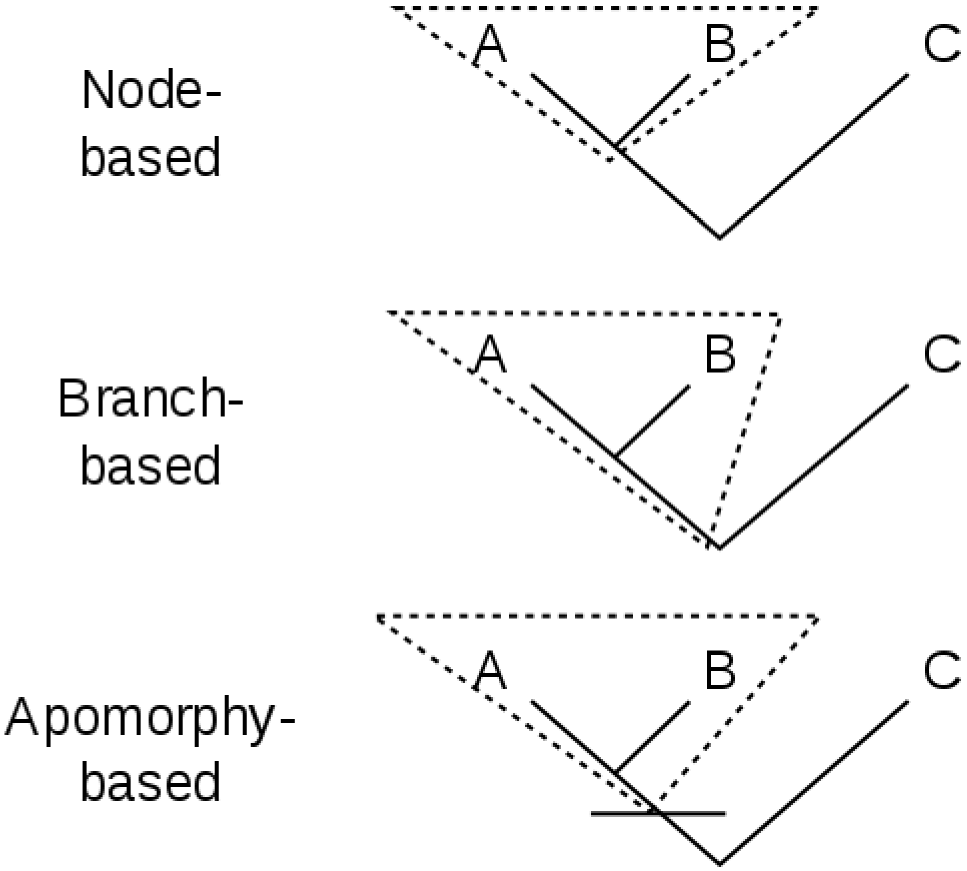


Clade types

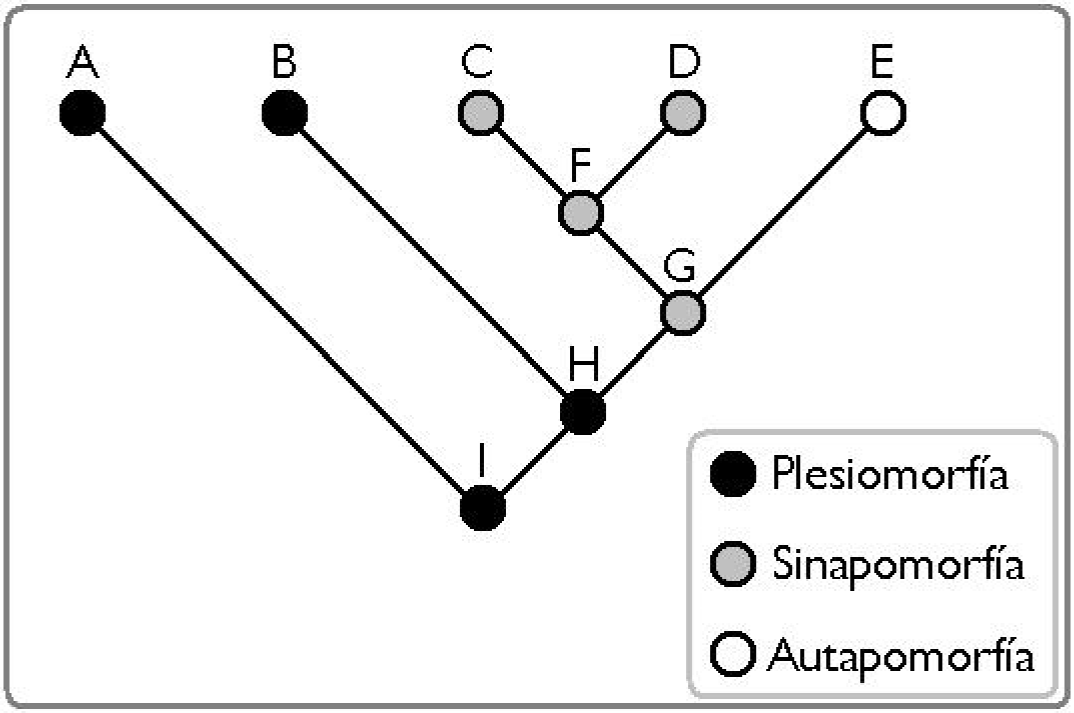
Molecular versus morphological data

The characteristics used to create a cladogram can be roughly categorized as either morphological (synapsid skull, warm blooded, notochord, unicellular, etc.) or molecular (DNA, RNA, or other genetic information). Prior to the advent of DNA sequencing, all cladistic analysis used morphological data.

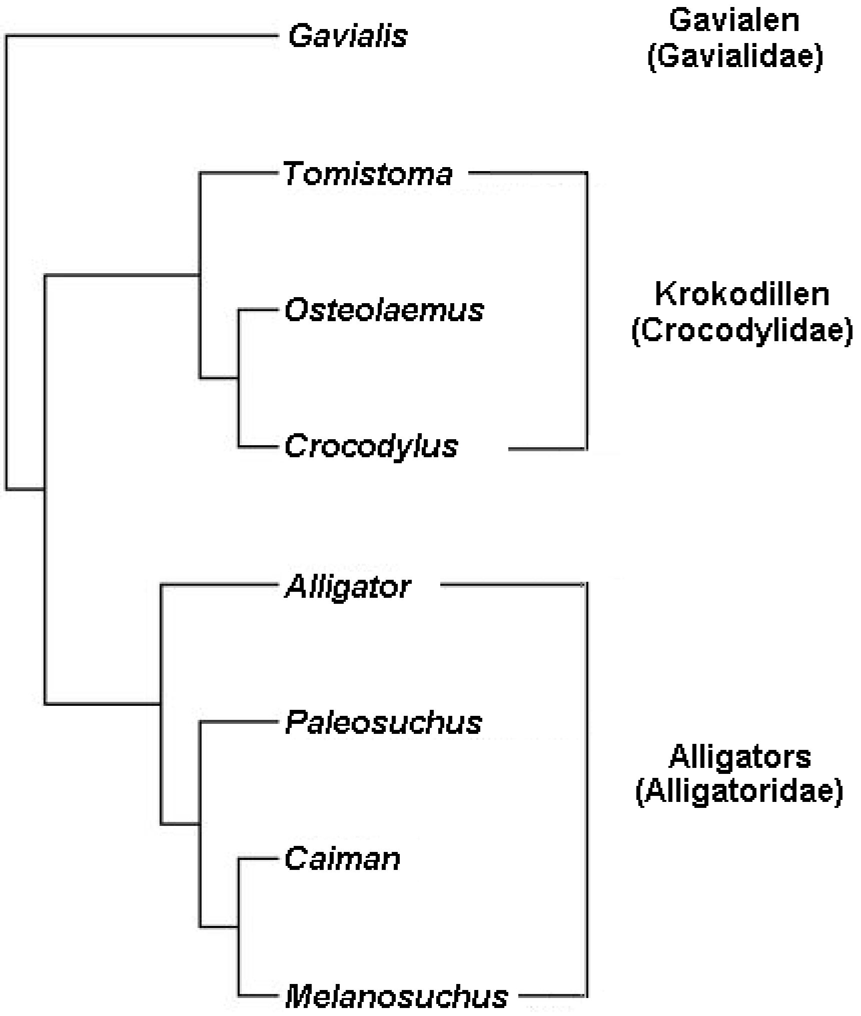
As DNA sequencing has become cheaper and easier, molecular systematics has become a more and more popular way to reconstruct phylogenies. Using a parsimony criterion is only one of several methods to infer a phylogeny from molecular data; maximum likelihood and Bayesian inference, which incorporate explicit models of sequence evolution, are non-Hennigian ways to evaluate sequence data. Another powerful method of reconstructing phylogenies is the use of genomic retrotransposon markers, which are thought to be less prone to the problem of reversion that plagues sequence data. They are also generally assumed to have a low incidence of homoplasies because it was once thought that their integration into the genome was entirely random; this seems at least sometimes not to be the case, however.



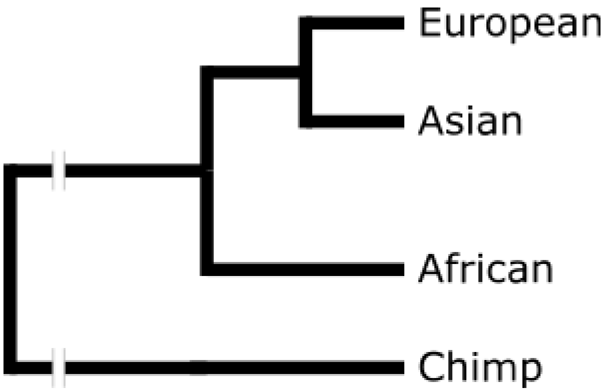
Clade types



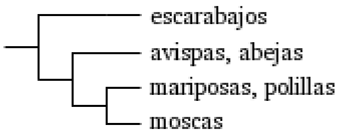
Clado



Cladogram Crocodilia...

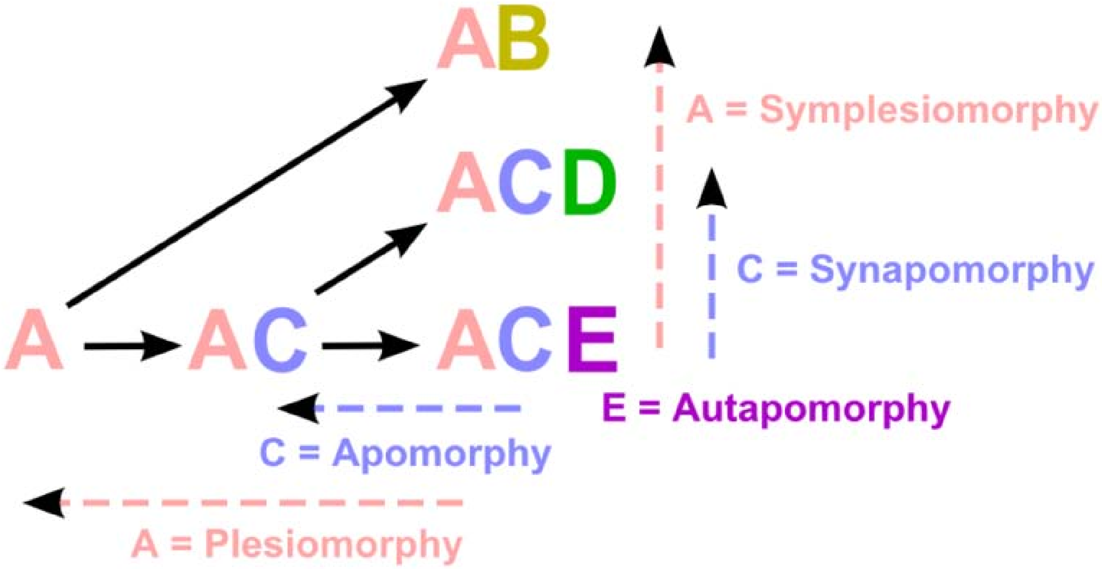


Cladogram of human p...



Cladogram-example1-e...

Ideally, morphological, molecular, and possibly other phylogenies should be combined into an analysis of *total evidence*: All have different intrinsic sources of error. For example, character convergence (homoplasy) is much more common in morphological data than in molecular sequence data, but character reversions that are unrecognizable as such are more common in the latter. Morphological homoplasies can usually be recognized as such if character states are defined with enough attention to detail.



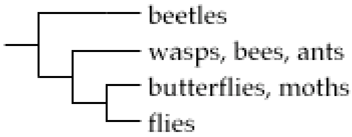
Apomorphy in cladistics

Plesiomorphies and synapomorphies

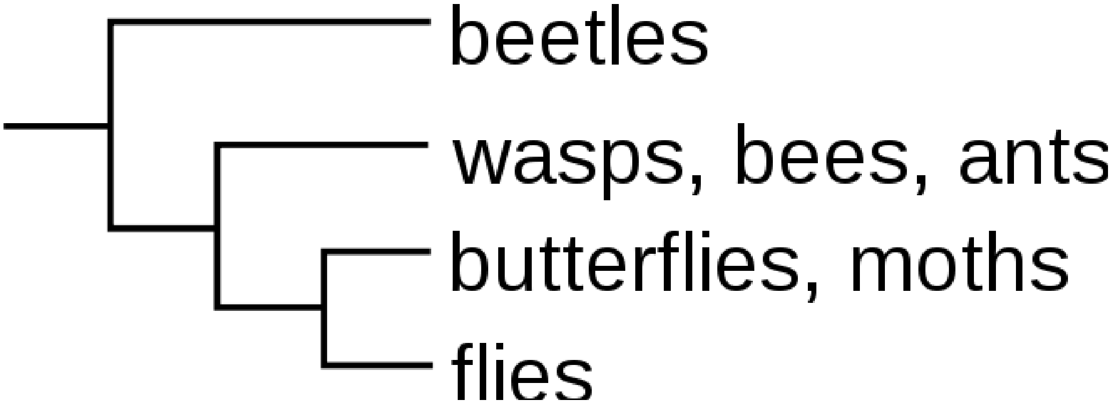
The researcher must decide which character states were present *before* the last common ancestor of the species group (*plesiomorphies*) and which were present *in* the last common ancestor (*synapomorphies*) and does so by comparison to one or more *outgroups*. The choice of an outgroup is a crucial step in cladistic analysis because different outgroups can produce trees with profoundly different topologies. Note that only synapomorphies are of use in characterizing clades.

Avoid homoplasies

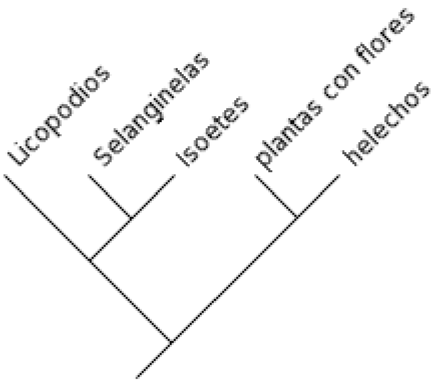
A homoplasy is a character that is shared by multiple species due to some cause *other* than common ancestry. The two main types of homoplasy are convergence (appearance of the same character in at least two distinct lineages) and reversion (the return to an ancestral character). Use of homoplasies when building a cladogram is sometimes unavoidable but is to be avoided when possible.



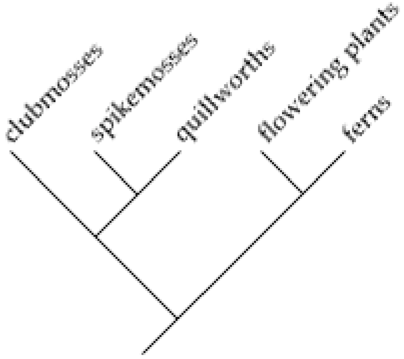
Cladogram-example1



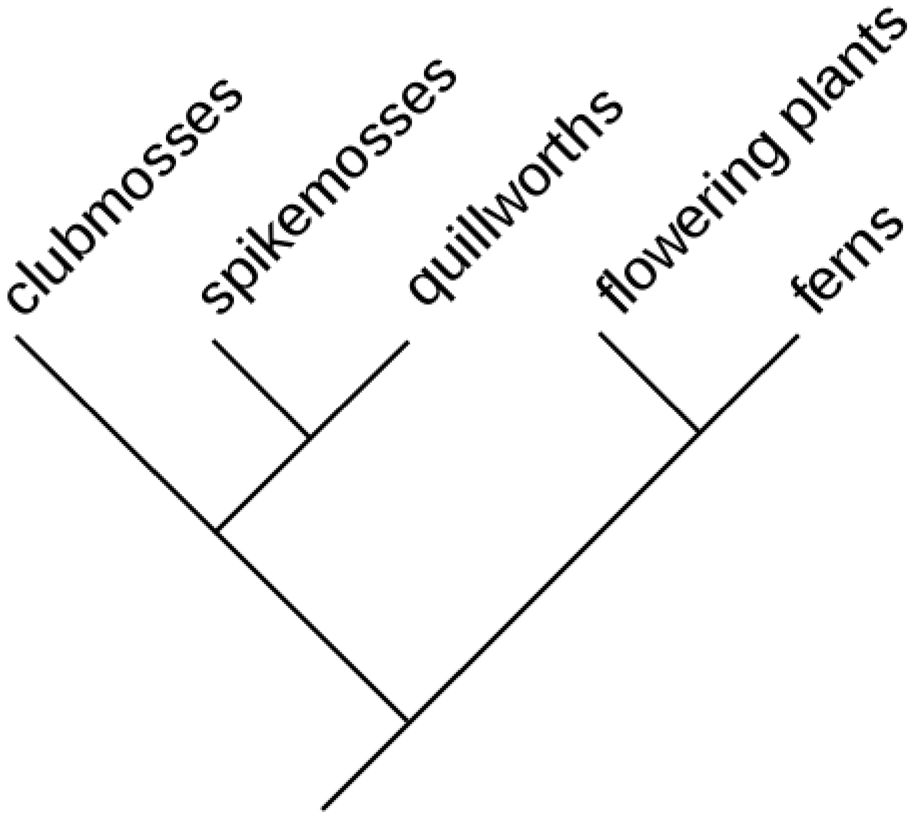
Cladogram-example1.svg



Cladogram-example2-e...

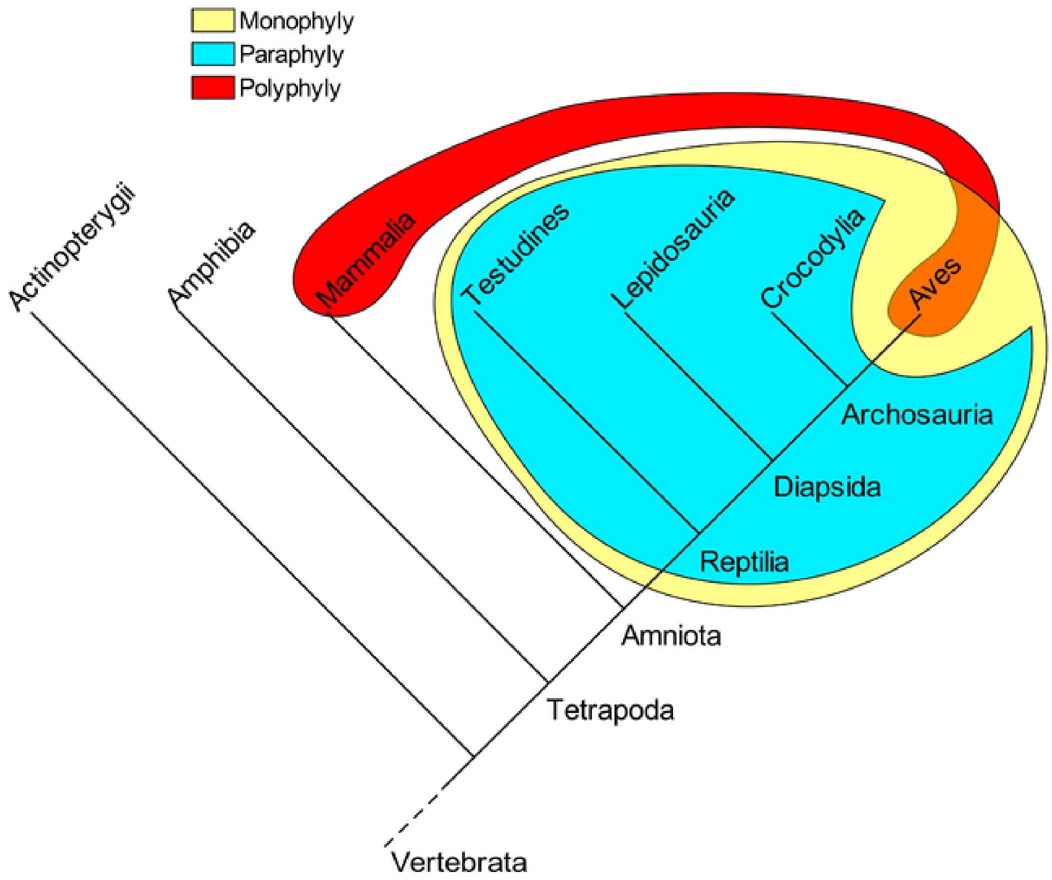


Cladogram-example2



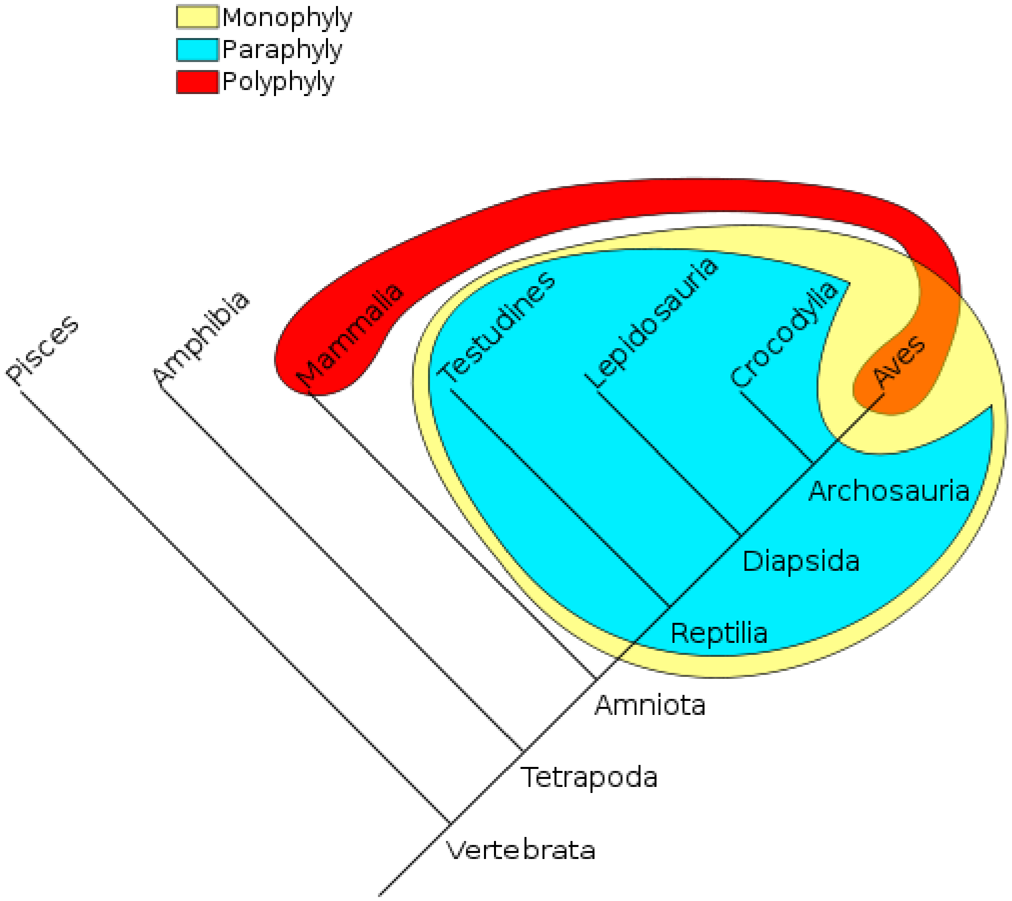
Cladogram-example2

A well known example of homoplasy due to convergent evolution would be the character, "presence of wings". Though the wings of birds, bats, and insects serve the same function, each evolved independently, as can be seen by their anatomy. If a bird, bat, and a winged insect were scored for the character, "presence of wings", a homoplasy would be introduced into the dataset, and this would confound the analysis, possibly resulting in a false evolutionary scenario.

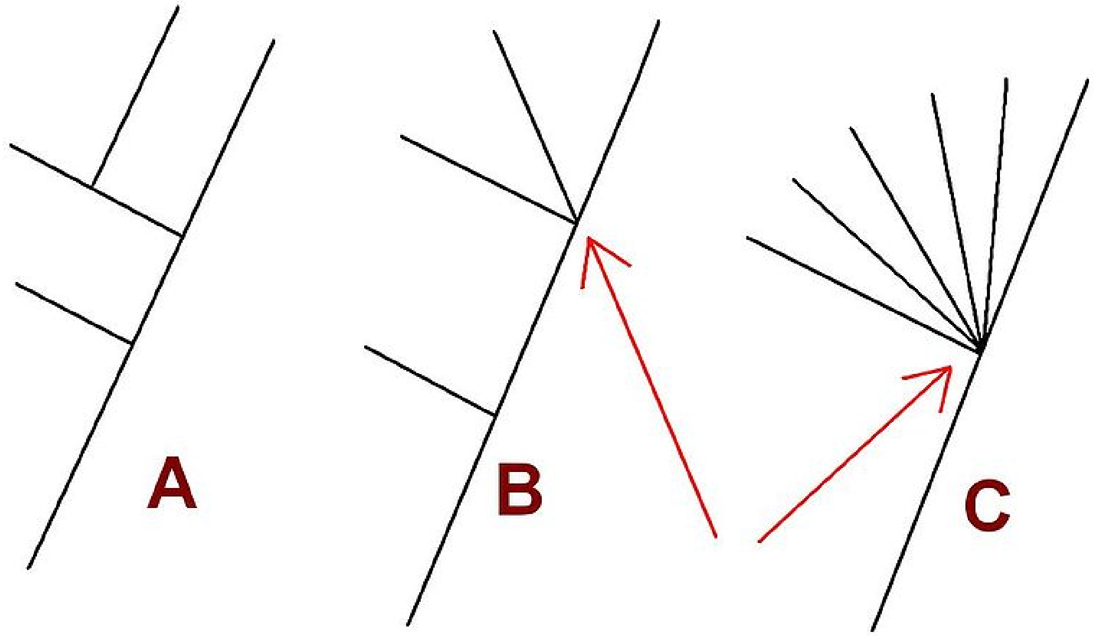


Partial Evolutionary Tree of the Vertebrates

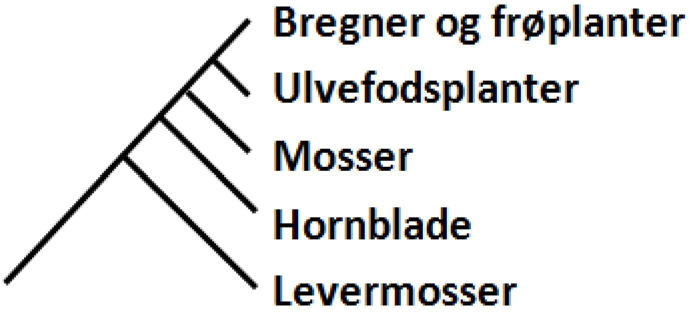
Phylogenetic-Groups-...



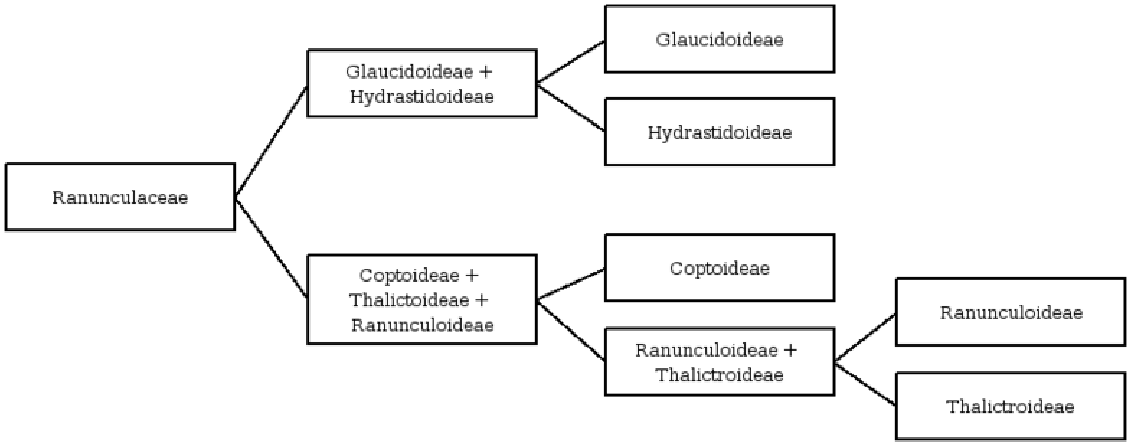
Phylogenetic-Groups



Polytomies

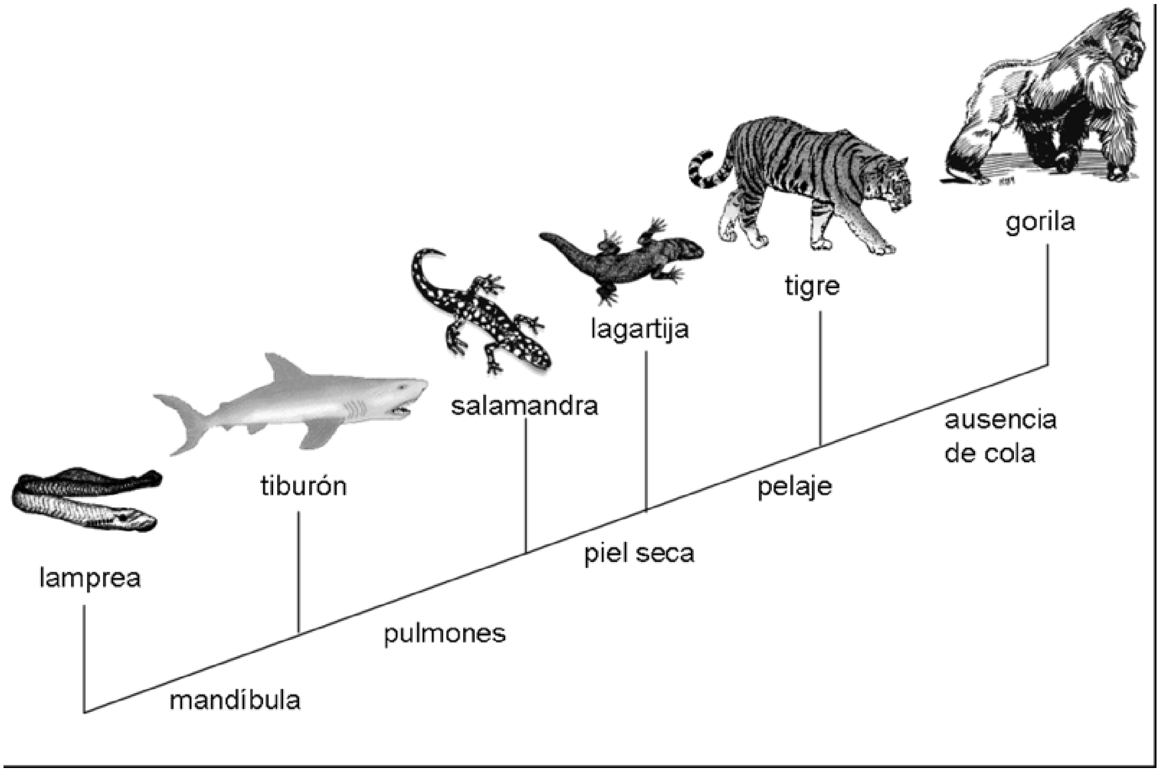


PrimitivePlanterClad...

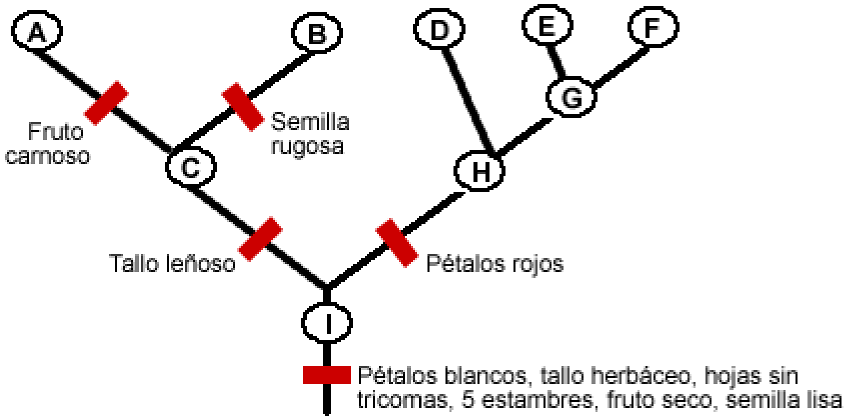

Ranunculaceae
Glaucidoideae + Hydrastidoideae
Glaucidoideae
Hydrastidoideae
Coptoideae + Thalictoideae + Ranunculoideae
Coptoideae
Ranunculoideae + Thalictroideae
Ranunculoideae
Thalictroideae

Ranunculaceae APGII

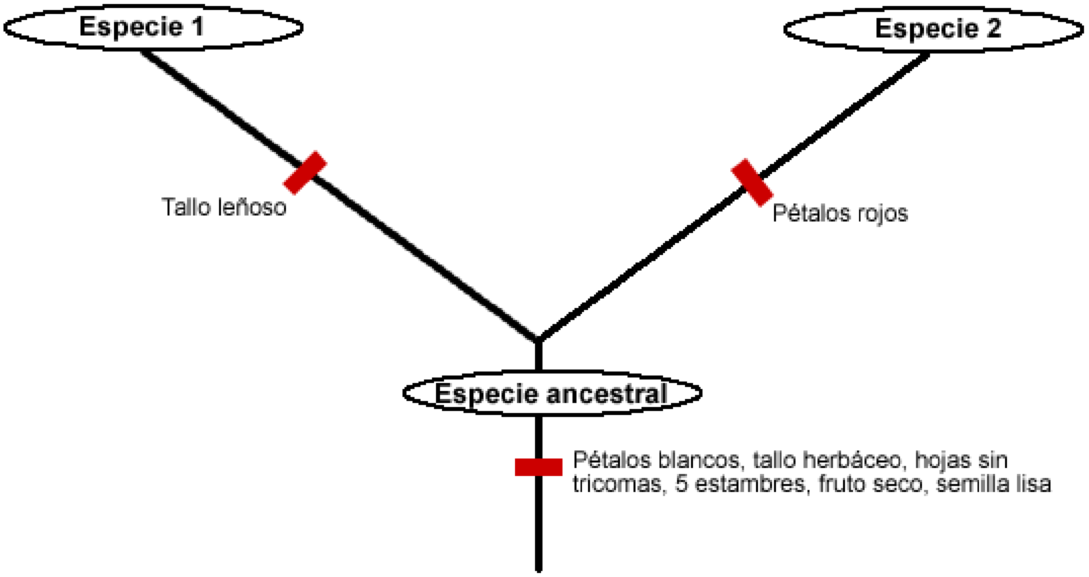
Homoplasies can often be avoided outright in morphological datasets by defining characters more precisely and increasing their number. When analyzing "supertrees" (datasets incorporating as many taxa of a suspected clade as possible), it may become unavoidable to introduce character definitions that are imprecise, as otherwise the characters might not apply at all to a large number of taxa; to continue with the "wings" example, the presence of wings would hardly be a useful character if attempting a phylogeny of all Metazoa, as most of these don't have wings at all. Cautious choice and definition of characters thus is another important element in cladistic analyses. With a faulty outgroup or character set, no method of evaluation is likely to produce a phylogeny representing the evolutionary reality.



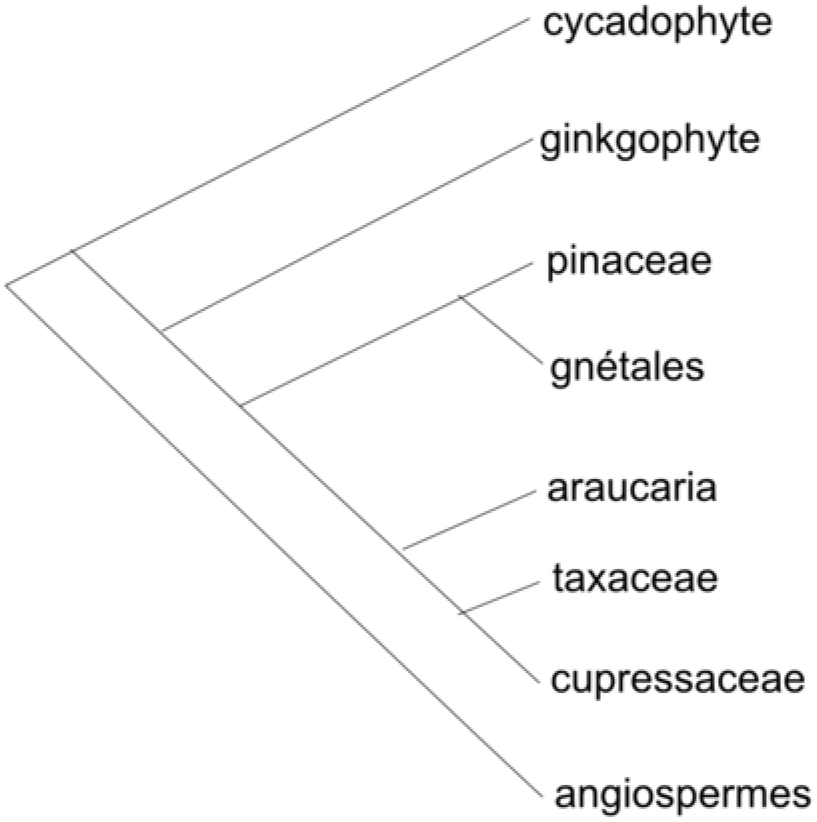
Cladograma


A
B
D
E
F
G
C
H
I
Fruto carnoso
Semilla rugosa
Tallo leñoso
Pétalos rojos
Pétalos blancos, tallo herbáceo, hojas sin tricomas, 5 estambres, fruto seco, semilla lisa

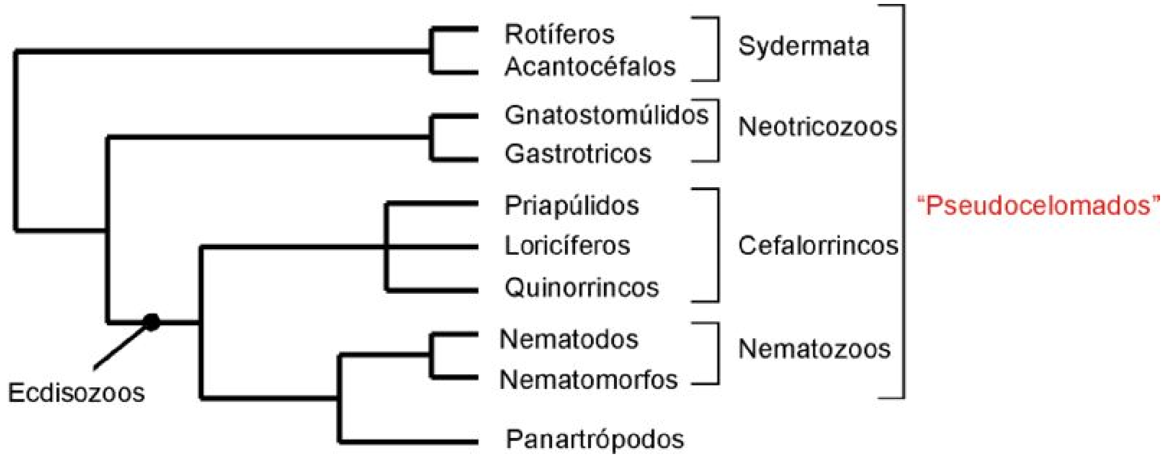
Cladogramanotansimpl...


Especie 1
Especie 2
Tallo leñoso
Pétalos rojos
Especie ancestral
Pétalos blancos, tallo herbáceo, hojas sin tricomas, 5 estambres, fruto seco, semilla lisa

Cladogramasimple

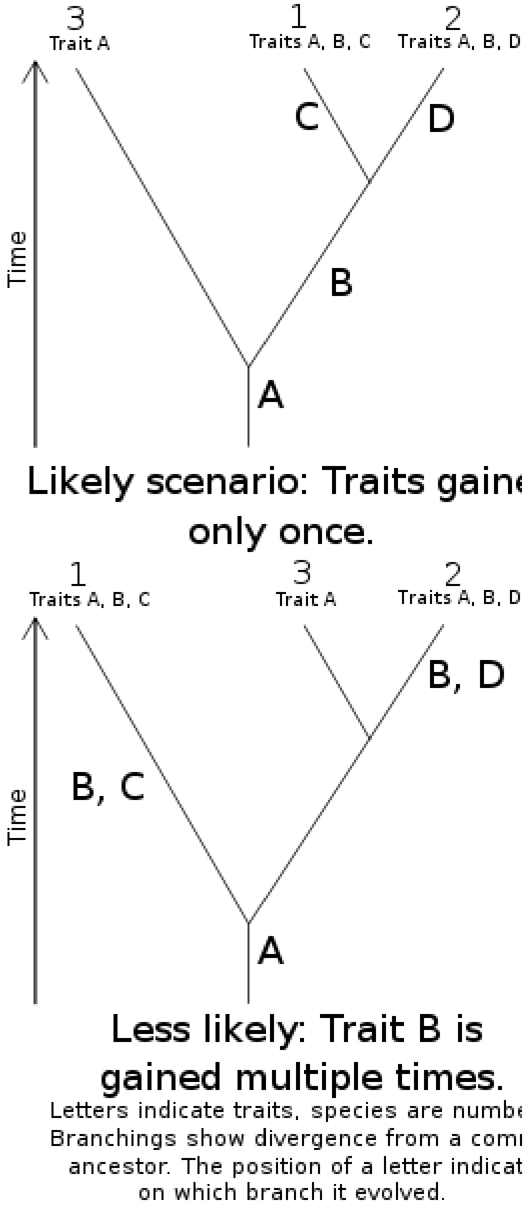


Cladogramme spermato...



Filogenia Pseudocelo...

Step 2: consider possible cladograms


3
Trait A
1
Traits A, B, C
2
Traits A, B, D
C
D
Time
B
A
Likely scenario: Traits gaine
only once.
1
Traits A, B, C
3
Trait A
2
Traits A, B, D
B, D
B, C
Time
A
Less likely: Trait B is
gained multiple times.
Letters indicate traits, species are numbe
Branchings show divergence from a comı
ancestor. The position of a letter indicat
on which branch it evolved.

Simple cladistics

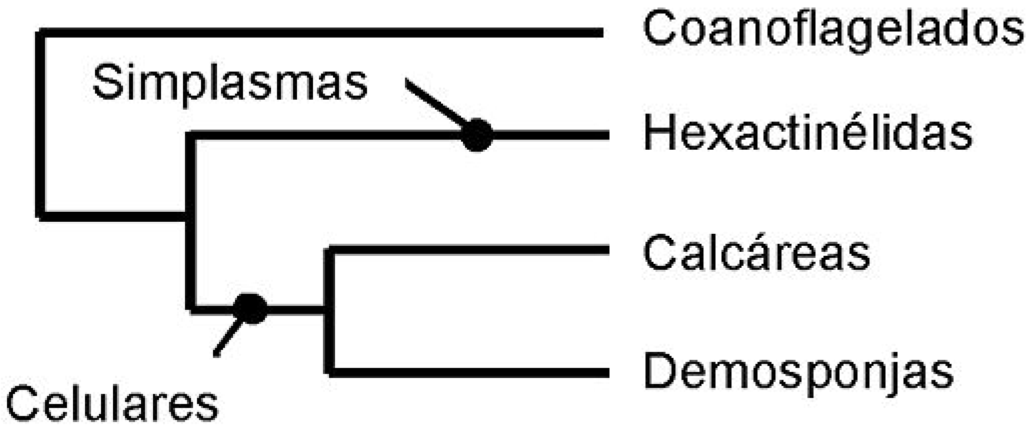
When there are just a few species being organized, it is possible to do this step manually, but most cases require a computer program. There are scores of computer programs available to support cladistics.

Because the total number of possible cladograms grows exponentially with the number of species, it is impractical for a computer program to evaluate every individual cladogram. A typical cladistic program begins by using heuristic techniques to identify a small number of candidate cladograms. Many cladistic programs then continue the search with the following repetitive steps:

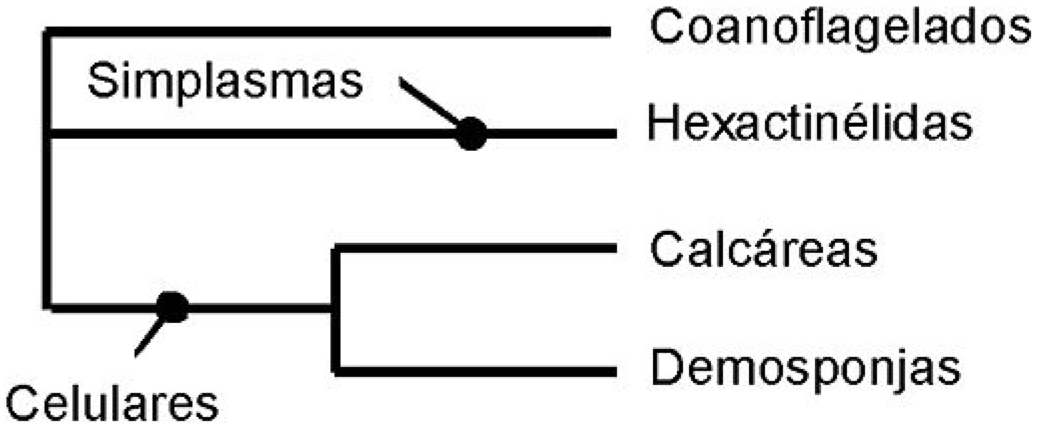
1. Evaluate the candidate cladograms by comparing them to the characteristic data

2. Identify the best candidates that are most consistent with the characteristic data
3. Create additional candidates by creating several variants of each of the best candidates from the prior step
4. Use heuristics to create several new candidate cladograms unrelated to the prior candidates
5. Repeat these steps until the cladograms stop getting better

Computer programs that generate cladograms use algorithms that are very computationally intensive, because the cladogram problem is NP-hard.

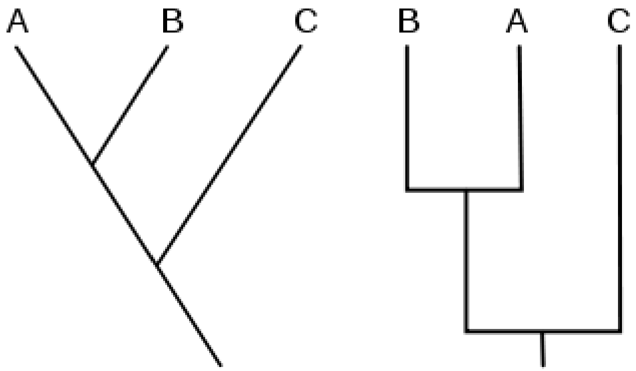


Monofilia de los Poríferos

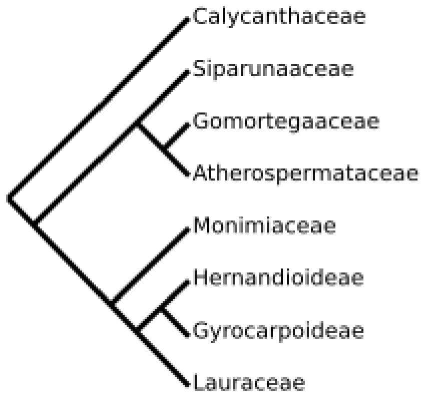


Polifilia de los Poríferos

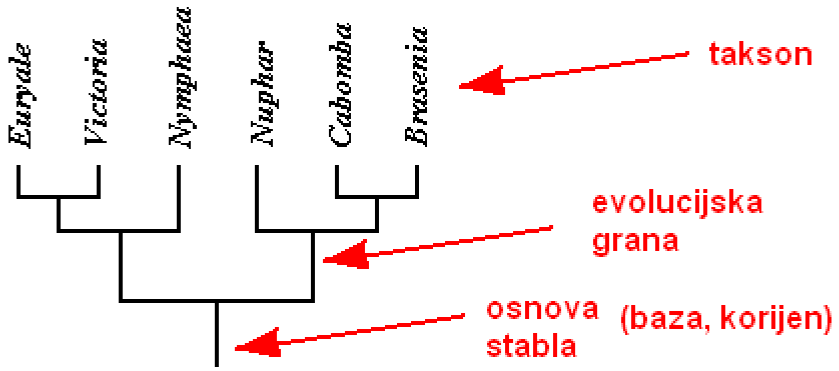
Filogenia porifera

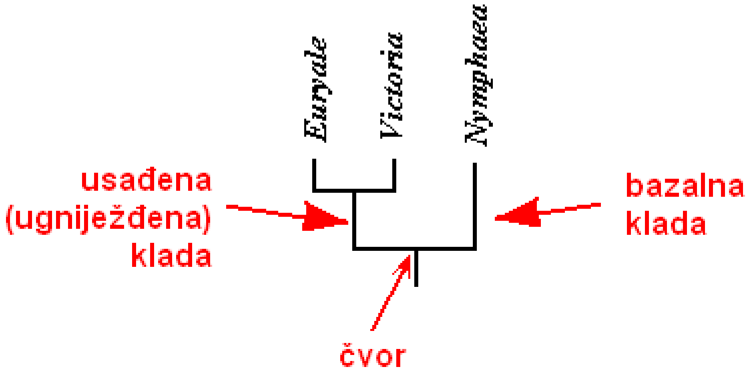


Identical cladograms...



Kladogram laurales





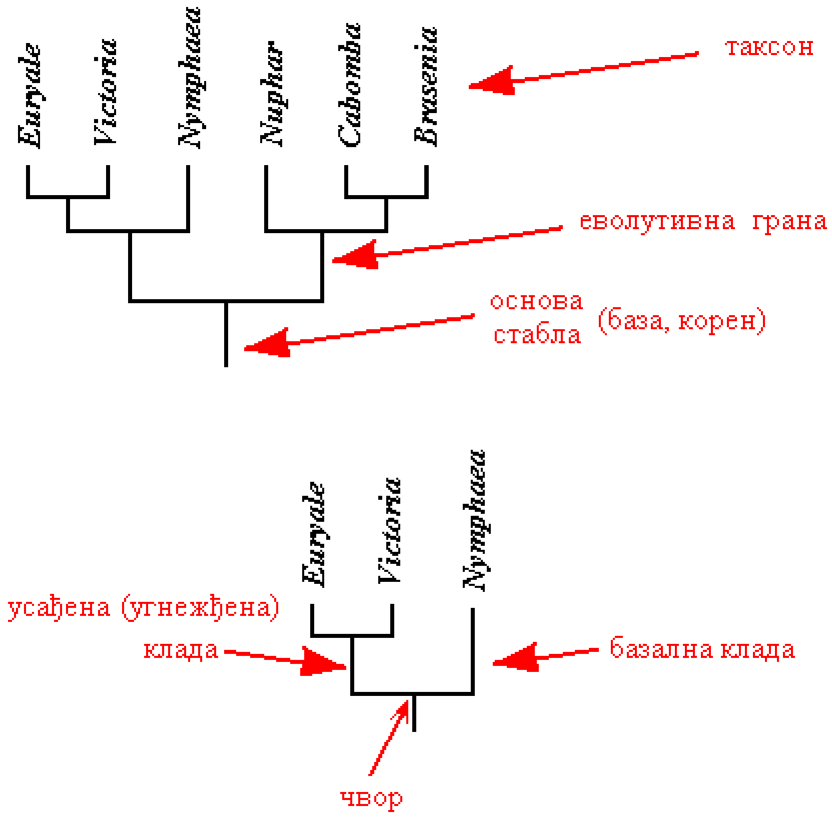
Kladogram objasnjenj...

Step 3: select best cladogram

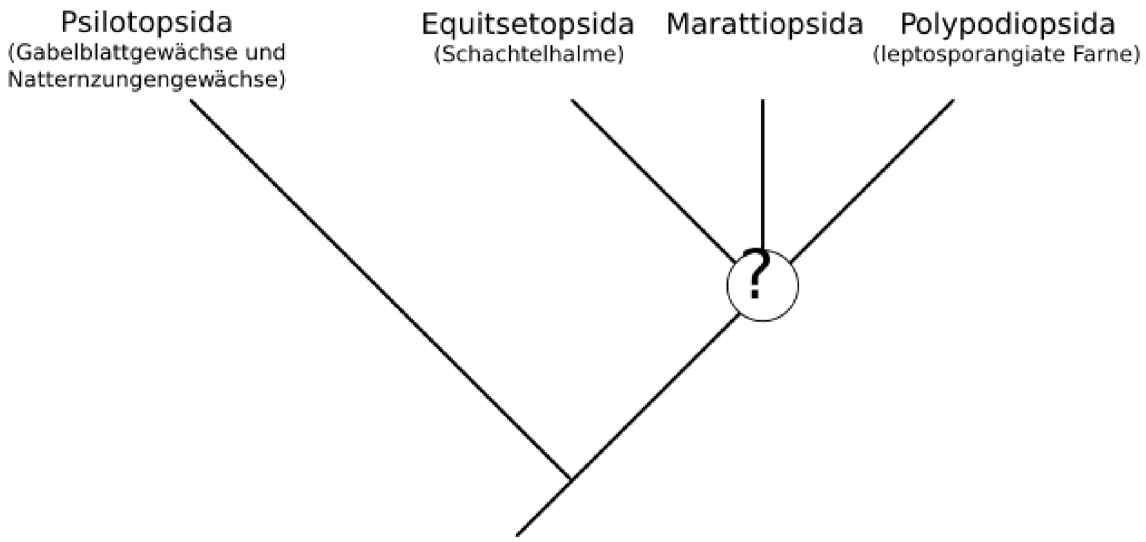
There are several algorithms available to identify the "best" cladogram. Most algorithms use a metric to measure how consistent a candidate cladogram is with the data. Most cladogram algorithms use the mathematical techniques of optimization and minimization.

In general, cladogram generation algorithms must be implemented as computer programs, although some algorithms can be performed manually when the data sets are trivial (for example, just a few species and a couple of characteristics).

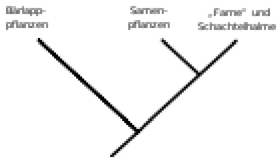
Some algorithms are useful only when the characteristic data are molecular (DNA, RNA); other algorithms are useful only when the characteristic data are morphological. Other algorithms can be used when the characteristic data includes both molecular and morphological data.



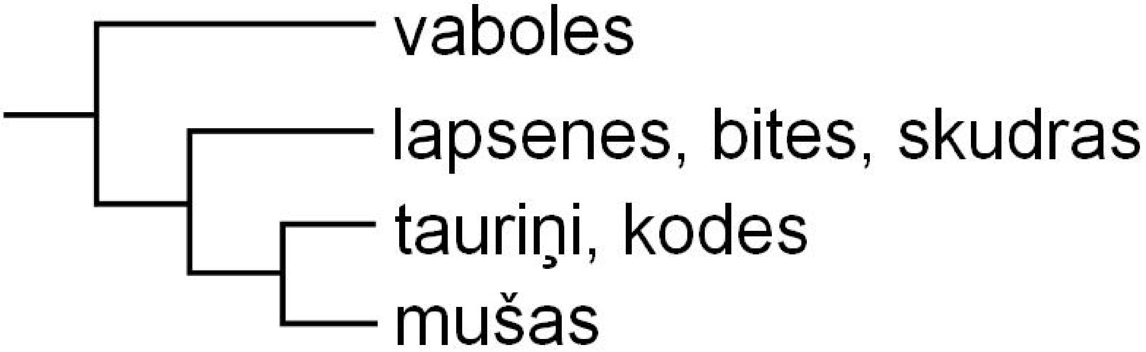
Kladogram objasnjenj...



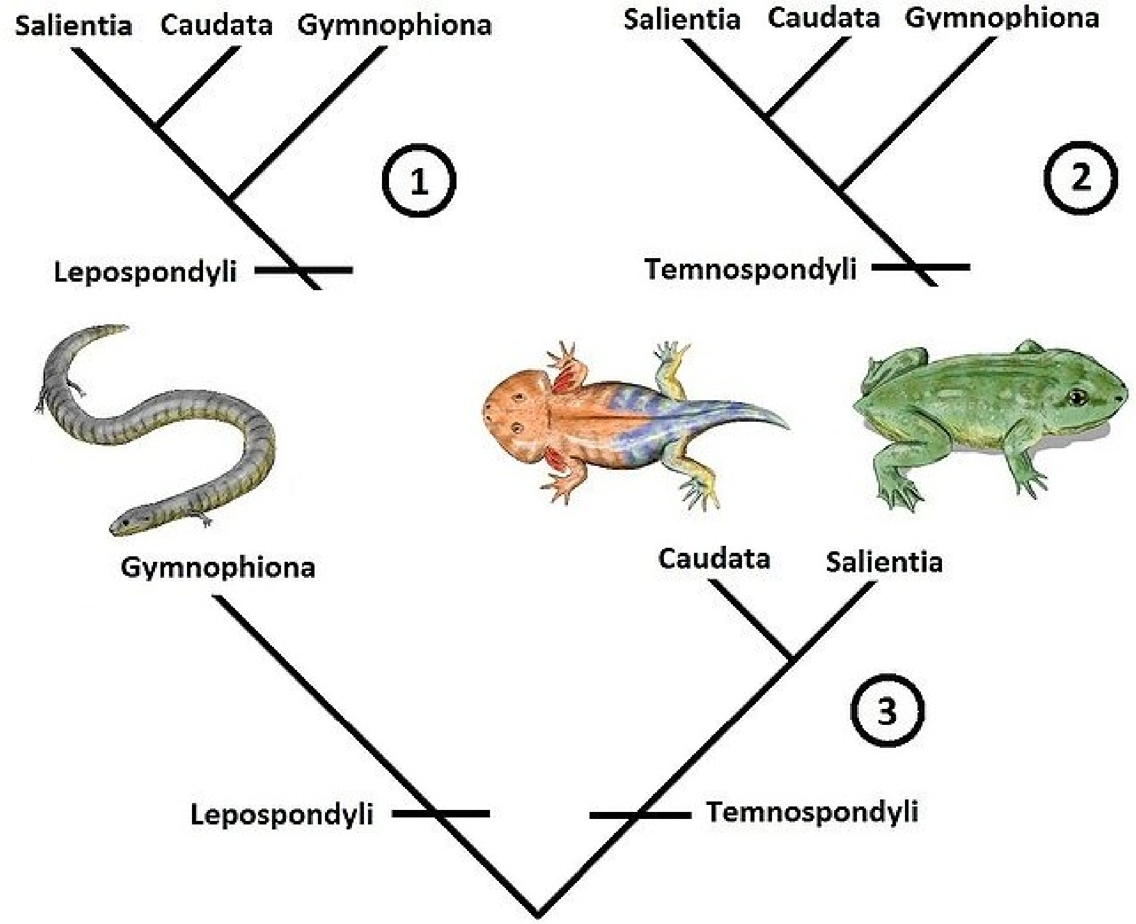
Kladogramm Farne+Sch...



Kladogramm Gefäßpf...



Kladogramma

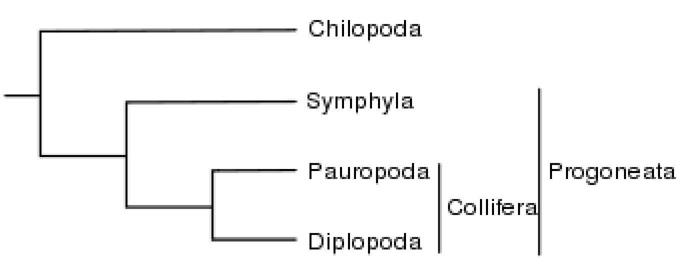


Lissamphibian phylog...

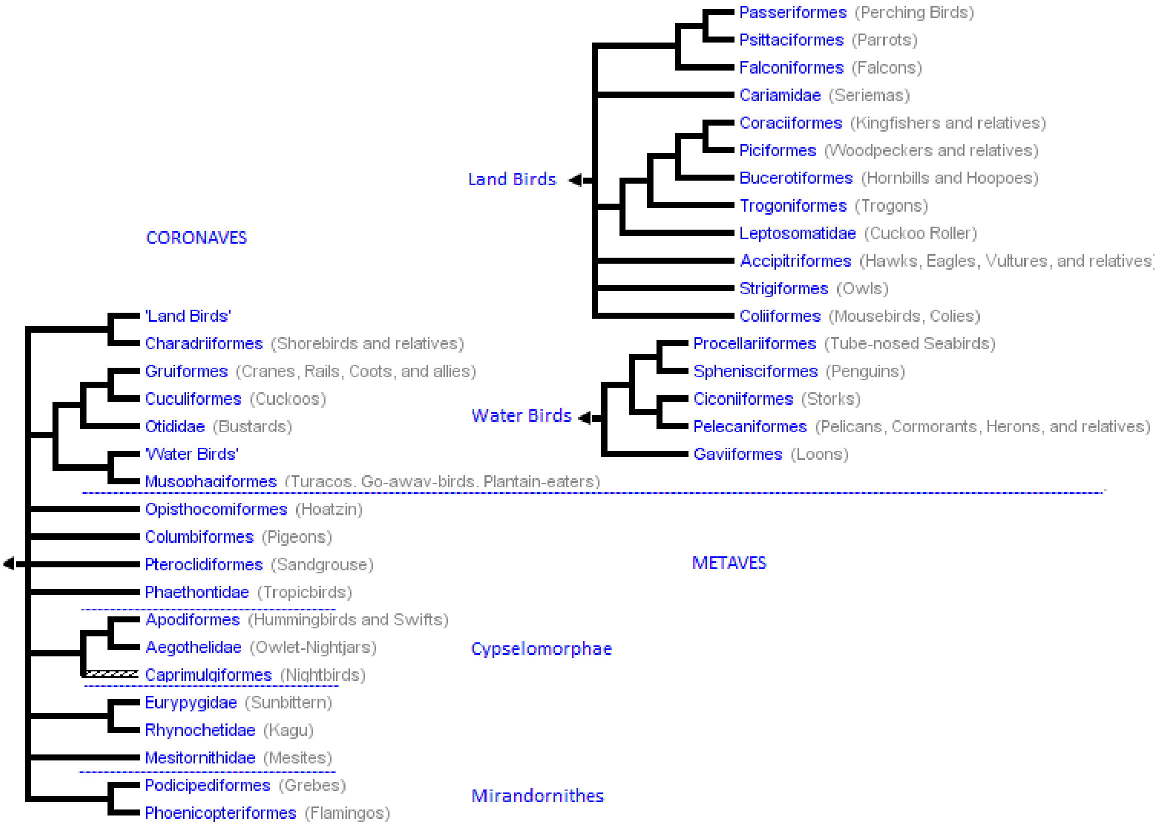
Algorithms for cladograms include least squares, neighbor-joining, parsimony, maximum likelihood, and Bayesian inference.

Biologists sometimes use the term parsimony for a specific kind of cladogram generation algorithm and sometimes as an umbrella term for all cladogram algorithms.

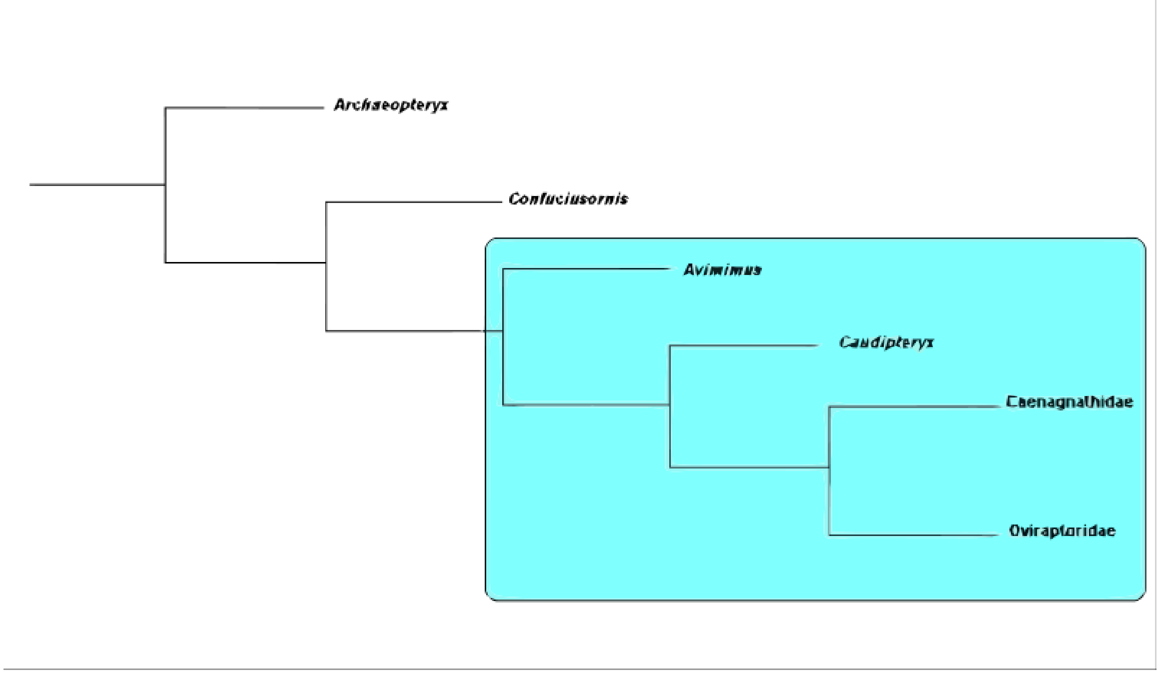
Algorithms that perform optimization tasks (such as building cladograms) can be sensitive to the order in which the input data (the list of species and their characteristics) is presented. Inputting the data in various orders can cause the same algorithm to produce different "best" cladograms. In these situations, the user should input the data in various orders and compare the results.



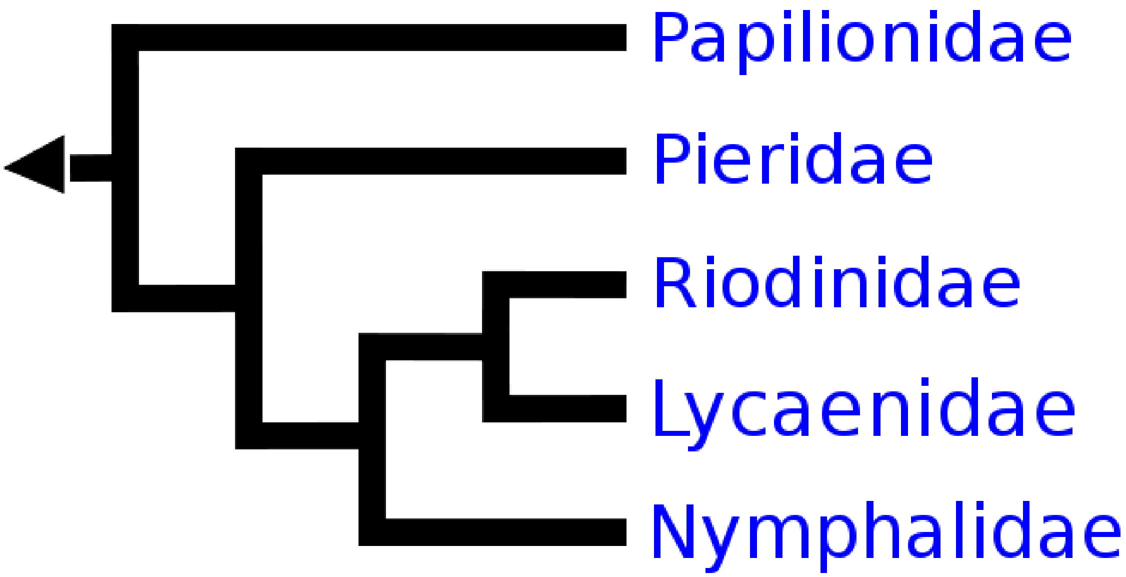
Myriapoda phylogeny


Passeriformes (Perching Birds)
Psittaciformes (Parrots)
Falconiformes (Falcons)
Cariamidae (Seriemas)
Coraciiformes (Kingfishers and relatives)
Piciformes (Woodpeckers and relatives)
Bucerotiformes (Hornbills and Hoopoes)
Trogoniformes (Trogons)
Leptosomatidae (Cuckoo Roller)
Accipitriformes (Hawks, Eagles, Vultures, and relatives)
Strigiformes (Owls)
Coliiformes (Mousebirds, Colies)
Land Birds
CORONAVES
'Land Birds'
Charadriiformes (Shorebirds and relatives)
Gruiformes (Cranes, Rails, Coots, and allies)
Cuculiformes (Cuckoos)
Otididae (Bustards)
'Water Birds'
Musophagiformes (Turacos, Go-away-birds, Plantain-eaters)
Procellariiformes (Tube-nosed Seabirds)
Sphenisciformes (Penguins)
Ciconiiformes (Storks)
Pelecaniformes (Pelicans, Cormorants, Herons, and relatives)
Gaviiformes (Loons)
Water Birds
Opisthocomiformes (Hoatzin)
Columbiformes (Pigeons)
Pteroclidiformes (Sandgrouse)
Phaethontidae (Tropicbirds)
METAVES
Apodiformes (Hummingbirds and Swifts)
Aegothelidae (Owlet-Nightjars)
Caprimulgiformes (Nightbirds)
Cypselomorphae
Eurypygidae (Sunbittern)
Rhynochetidae (Kagu)
Mesitornithidae (Mesites)
Podicipediformes (Grebes)
Phoenicopteriformes (Flamingos)
Mirandornithes

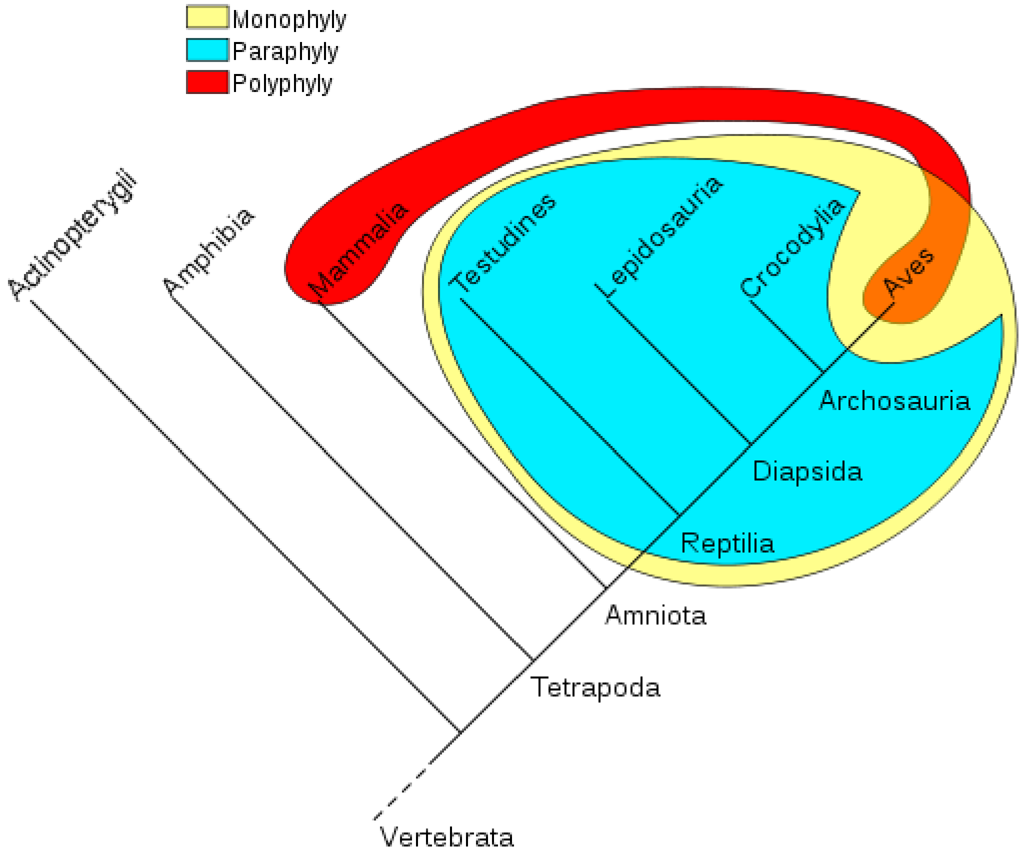
Neoaves Alternative ...



Outgroup



Papilionoidea

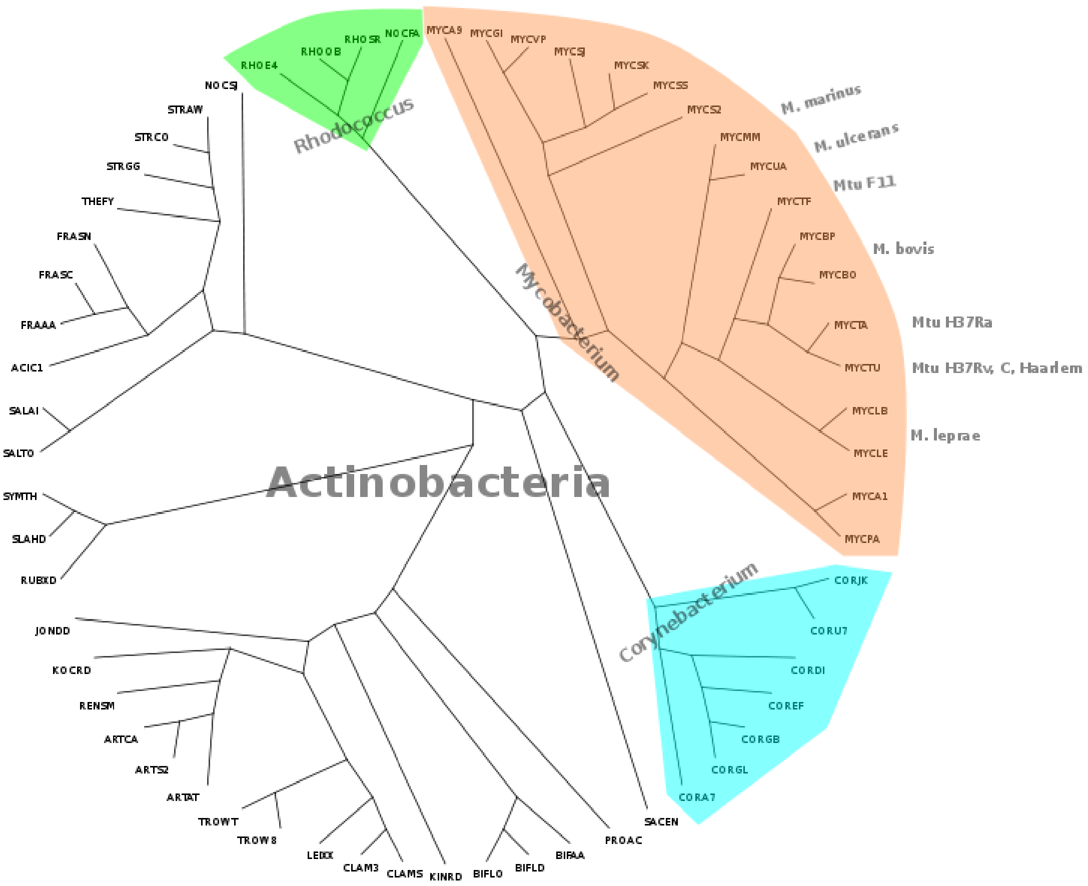


Partial Evolutionary Tree of the Vertebrates

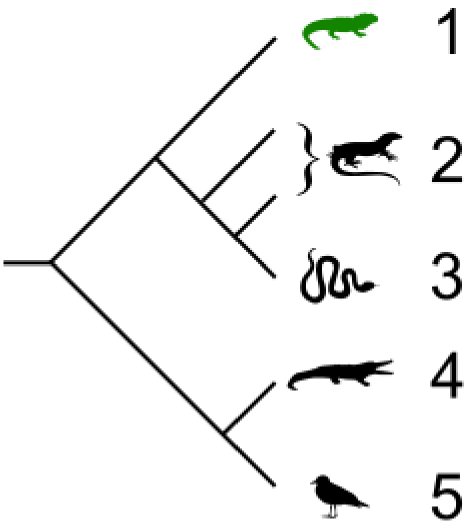
Phylogenetic-Groups-...

Using different algorithms on a single data set can sometimes yield different "best" cladograms, because each algorithm may have a unique definition of what is "best".

Because of the astronomical number of possible cladograms, algorithms cannot guarantee that the solution is the overall best solution. A nonoptimal cladogram will be selected if the program settles on a local minimum rather than the desired global minimum. To help solve this problem, many cladogram algorithms use a simulated annealing approach to increase the likelihood that the selected cladogram is the optimal one.



TreeActinobacteria



Tuatara cladogram

Chapter- 6

Biological Classification

The hierarchy of biological classification's eight major taxonomic ranks, which is an example of definition by genus and differentia. Intermediate minor rankings are not shown.

Biological classification, or *scientific classification in biology*, is a method by which biologists group and categorize organisms by biological type, such as genus or species. Biological classification is a form of scientific taxonomy.

Modern biological classification has its root in the work of Carolus Linnaeus, who grouped species according to shared physical characteristics. These groupings have since been revised to improve consistency with the Darwinian principle of common descent. Molecular phylogenetics, which uses DNA sequences as data, has driven many recent revisions and is likely to continue to do so. Biological classification belongs to the science of biological systematics.

Taxonomic ranks

In biological classification, **rank** is the level (the relative position) in a hierarchy. Sometimes (but only rarely) the term "taxonomic category" is used instead of "rank". There are 7 main ranks defined by the international nomenclature codes: Kingdom, phylum/division, class, order, family, genus, species. "Domain", a level above kingdom, has become popular in recent years, but has not been accepted into the codes.

The most basic rank is that of species, the next most important is genus, and then family.

The *International Code of Zoological Nomenclature* defines rank, in the nomenclatural sense, as:

The level, for nomenclatural purposes, of a taxon in a taxonomic hierarchy (e.g. all families are for nomenclatural purposes at the same rank, which lies between superfamily and subfamily). The ranks of the family group, the genus group, and the species group at which nominal taxa may be established are stated in Articles 10.3, 10.4, 35.1, 42.1 and 45.1.

There are slightly different ranks for zoology and for botany, including subdivisions such as tribe.

Early systems

Ancient through medieval times

Current systems of classifying forms of life descend from the thought presented by the Greek philosopher Aristotle, who published in his metaphysical works the first known classification of everything whatsoever, or "being". This is the scheme that gave such words as 'substance', 'species' and 'genus' and was retained in modified and less general form by Linnaeus.

Aristotle also studied animals and classified them according to method of reproduction, as did Linnaeus later with plants. Aristotle's animal classification was eventually made obsolete by additional knowledge and forgotten.

The philosophical classification is in brief as follows: Primary substance is the individual being; for example, Peter, Paul, etc. Secondary substance is a predicate that can properly or characteristically be said of a class of primary substances; for example, man of Peter, Paul, etc. The characteristic must not be merely in the individual; for example, being skilled in grammar. Grammatical skill leaves most of Peter out and therefore is not characteristic of him. Similarly man (all of mankind) is not in Peter; rather, he is in man.

Species is the secondary substance that is most proper to its individuals. The most characteristic thing that can be said of Peter is that Peter is a man. An identity is being postulated: "man" is equal to all its individuals and only those individuals. Members of a species differ only in number but are totally the same type.

Genus is a secondary substance less characteristic of and more general than the species; for example, man is an animal, but not all animals are men. It is clear that a genus contains species. There is no limit to the number of Aristotelian genera that might be found to contain the species. Aristotle does not structure the genera into phylum, class, etc., as the Linnaean classification does.

The secondary substance that distinguishes one species from another within a genus is the specific difference. Man can thus be comprehended as the sum of specific differences (the "differentiae" of biology) in less and less general categories. This sum is the definition; for example, man is an animate, sensate, rational substance. The most characteristic definition contains the species and the next most general genus: man is a rational animal. Definition is thus based on the unity problem: the species is but one yet has many differentiae.

The very top genera are the categories. There are ten: one of substance and nine of "accidents", universals that must be "in" a substance. Substances exist by themselves; accidents are only in them: quantity, quality, etc. There is no higher category, "being", because of the following problem, which was only solved in the Middle Ages by Thomas Aquinas: a specific difference is not characteristic of its genus. If man is a rational animal, then rationality is not a property of animals. Substance therefore cannot be a *kind* of being because it can have no specific difference, which would have to be *non*-being.

The problem of being occupied the attention of scholastics during the time of the Middle Ages. The solution of St. Thomas, termed the analogy of being, established the field of ontology, which received the better part of the publicity and also drew the line between philosophy and experimental science. The latter rose in the Renaissance from practical technique. Linnaeus, a classical scholar, combined the two on the threshold of the neo-classicist revival now called the Age of Enlightenment.

Renaissance through Age of Reason

An important advance was made by the Swiss professor, Conrad von Gesner (1516–1565). Gesner's work was a critical compilation of life known at the time.

The exploration of parts of the New World by Europeans produced large numbers of new plants and animals that needed descriptions and classification. The old systems made it difficult to study and locate all these new specimens within a collection and often the same plants or animals were given different names simply because there were too many species to keep track of. A system was needed that could group these specimens together so they could be found; the binomial system was developed based on morphology with groups having similar appearances. In the latter part of the 16th century and the beginning of the 17th, careful study of animals commenced, which, directed first to familiar kinds, was gradually extended until it formed a sufficient body of knowledge to serve as an anatomical basis for classification. Advances in using this knowledge to classify living beings bear a debt to the research of medical anatomists, such as Fabricius (1537–1619), Petrus Severinus (1580–1656), William Harvey (1578–1657), and Edward Tyson (1649–1708). Advances in classification due to the work of entomologists and the first microscopists is due to the research of people like Marcello Malpighi (1628–1694), Jan Swammerdam (1637–1680), and Robert Hooke (1635–1702). Lord Monboddo (1714–1799) was one of the early abstract thinkers whose works illustrate knowledge of species relationships and who foreshadowed the theory of evolution..

Early methodists

Since late in the 15th century, a number of authors had become concerned with what they called *methodus,* (method). By method authors mean an arrangement of minerals, plants, and animals according to the principles of logical division. The term *Methodists* was coined by Carolus Linnaeus in his *Bibliotheca Botanica* to denote the authors who care about the principles of classification (in contrast to the mere *collectors* who are concerned primarily with the description of plants paying little or no attention to their arrangement into genera, etc.). Important early Methodists were Italian philosopher, physician, and botanist Andrea Caesalpino, English naturalist John Ray, German physician and botanist Augustus Quirinus Rivinus, and French physician, botanist, and traveller Joseph Pitton de Tournefort.

Andrea Caesalpino (1519–1603) in his *De plantis libri XVI* (1583) proposed the first methodical arrangement of plants. On the basis of the structure of trunk and fructification he divided plants into fifteen "higher genera".

John Ray (1627–1705) was an English naturalist who published important works on plants, animals, and natural theology. The approach he took to the classification of plants in his Historia Plantarum was an important step towards modern taxonomy. Ray rejected the system of dichotomous division by which species were classified according to a pre-conceived, either/or type system, and instead classified plants according to similarities and differences that emerged from observation.

Both Caesalpino and Ray used traditional plant names and thus, the name of a plant did not reflect its taxonomic position (e.g. even though the apple and the peach belonged to different "higher genera" of John Ray's *methodus*, both retained their traditional names *Malus* and *Malus Persica* respectively). A further step was taken by Rivinus and Pitton de Tournefort who made genus a distinct rank within taxonomic hierarchy and introduced the practice of naming the plants according to their genera.

Augustus Quirinus Rivinus (1652–1723), in his classification of plants based on the characters of the flower, introduced the category of order (corresponding to the "higher" genera of John Ray and Andrea Caesalpino). He was the first to abolish the ancient division of plants into herbs and trees and insisted that the true method of division should be based on the parts of the fructification alone. Rivinus extensively used dichotomous keys to define both orders and genera. His method of naming plant species resembled that of Joseph Pitton de Tournefort. The names of all plants belonging to the same genus should begin with the same word (generic name). In the genera containing more than one species the first species was named with generic name only, while the second, etc. were named with a combination of the generic name and a modifier (*differentia specifica*).

Joseph Pitton de Tournefort (1656–1708) introduced an even more sophisticated hierarchy of class, section, genus, and species. He was the first to use consistently the uniformly composed species names that consisted of a generic name and a many-worded diagnostic phrase *differentia specifica.* Unlike Rivinus, he used *differentiae* with all species of polytypic genera.

Linnaean

Carolus Linnaeus' great work, the *Systema Naturæ* (1st ed. 1735), ran through twelve editions during his lifetime. In this work, nature was divided into three kingdoms: mineral, vegetable and animal. Linnaeus used five ranks: class, order, genus, species, and variety.

He abandoned long descriptive names of classes and orders and two-word generic names (e. g. *Trifolium repens*) still used by his immediate predecessors (Rivinus and Pitton de Tournefort) and replaced them with single-word names, provided genera with detailed diagnoses (*characteres naturales*), and reduced numerous varieties to their species, thus saving botany from the chaos of new forms produced by horticulturalists.

Linnaeus is best known for his introduction of the method still used to formulate the scientific name of every species. Before Linnaeus, long many-worded names (composed of a generic name and a *differentia specifica*) had been used, but as these names gave a description of the species, they were not fixed. In his *Philosophia Botanica* (1751) Linnaeus took every effort to improve the composition and reduce the length of the many-worded names by abolishing unnecessary rhetorics, introducing new descriptive terms and defining their meaning with an unprecedented precision. In the late 1740s Linnaeus began to use a parallel system of naming species with *nomina trivialia. Nomen triviale*, a trivial name, was a single- or two-word epithet placed on the margin of the page next to the many-worded "scientific" name. The only rules Linnaeus applied to them was that the trivial names should be short, unique within a given genus, and that they should not be changed. Linnaeus consistently applied *nomina trivialia* to the species of plants in *Species Plantarum* (1st edn. 1753) and to the species of animals in the 10th edition of *Systema Naturæ* (1758).

By consistently using these specific epithets, Linnaeus separated nomenclature from taxonomy. Even though the parallel use of *nomina trivialia* and many-worded descriptive names continued until late in the eighteenth century, it was gradually replaced by the practice of using shorter proper names combined of the generic name and the trivial name of the species. In the nineteenth century, this new practice was codified in the first Rules and Laws of Nomenclature, and the 1st edn. of *Species Plantarum* and the 10th edn. of *Systema Naturae* were chosen as starting points for the Botanical and Zoological Nomenclature respectively. This convention for naming species is referred to as binomial nomenclature.

Today, nomenclature is regulated by Nomenclature Codes, which allows names divided into taxonomic ranks.

Modern system

Whereas Linnaeus classified for ease of identification, it is now generally accepted that classification should reflect the Darwinian principle of common descent. Since the 1960s a trend called cladistic taxonomy (or cladistics or cladism) has emerged, arranging taxa in an evolutionary tree. If a taxon includes all the descendants of some ancestral form, it is called monophyletic. Groups that has descendant groups removed from it (e.g. dinosaurs, with birds as offspring group) are termed paraphyletic, while groups representing more than one branch from the tree of life (science) are called polyphyletic.

A new formal code of nomenclature, the *International Code of Phylogenetic Nomenclature*, or *PhyloCode* for short, is currently under development, intended to deal with names of clades. Linnaean ranks will be optional under the *PhyloCode*, which is intended to coexist with the current, rank-based codes.

Kingdoms and domains

From well before Linnaeus, plants and animals were considered separate Kingdoms. Linnaeus used this as the top rank, dividing the physical world into the plant, animal and mineral kingdoms. As advances in microscopy made classification of microorganisms possible, the number of kingdoms increased, five and six-kingdom systems being the most common.

Domains are a relatively new grouping. The three-domain system was first invented in 1990, but not generally accepted until later. The majority of biologists use the 5-Kingdom system, but a large minority use the 3 domain system. One main characteristic of the three-domain method is the separation of Archaea and Bacteria, previously grouped into the single kingdom Bacteria (a kingdom also sometimes called Monera). Consequently, the three domains of life are conceptualized as Archaea, Bacteria, and Eukaryota (comprising the nuclei-bearing eukaryotes). A small minority of scientists add Archaea as a sixth kingdom, but do not accept the domain method.

Thomas Cavalier-Smith, who has published extensively on the classification of protists, has recently proposed that the Neomura, the clade that groups together the Archaea and Eukarya, would have evolved from Bacteria, more precisely from Actinobacteria. His classification of 2004 treats the archaebacteria as part of a subkingdom of the Kingdom Bacteria, i.e. he rejects the three-domain system entirely.

Linnaeus 1735 2 kingdoms	**Haeckel 1866 3 kingdoms**	**Chatton 1925 2 empires**	**Copeland 1938 4 kingdoms**	**Whittaker 1969 5 kingdoms**	**Woese et al. 1977 6 kingdoms**	**Woese et al. 1990 3 domains**	**Cavalier-Smith 2004 6 kingdoms**
(not treated)	Protista	Prokaryota	Mychota	Monera	Eubacteria	Bacteria	Bacteria
					Archaebacteria	Archaea	
		Eukaryota	Protoctista	Protista	Protista	Eukarya	Protozoa
							Chromista
Vegetabilia	Plantae		Plantae	Plantae	Plantae		Plantae
			Protoctista	Fungi	Fungi		Fungi
Animalia	Animalia		Animalia	Animalia	Animalia		Animalia

Authorities (author citation)

The name of any taxon may be followed by the "authority" for the name, that is, the name of the author who first published a valid description of it. These names are frequently abbreviated: the abbreviation "L." is universally accepted for Linnaeus, and in botany there is a regulated list of standard abbreviations. The system for assigning authorities is slightly different in different branches of biology. However, it is standard that if a name or placement has been changed since the original description, the first authority's name is placed in parentheses and the authority for the new name or placement may be placed after it (usually only in botany and zoology).

Globally Unique Identifiers for names

There is a movement within the biodiversity informatics community to provide Globally Unique Identifiers in the form of Life Science Identifiers (LSID) for all biological names. This would allow authors to cite names unambiguously in electronic media and reduce the significance of errors in the spelling of names or the abbreviation of authority names. Three large nomenclatural databases (referred to as nomenclators) have already begun this process, these are Index Fungorum, International Plant Names Index and ZooBank. Other databases, that publish taxonomic rather than nomenclatural data, have also started using LSIDs to identify **taxa**. The key example of this is Catalogue of Life. The next step in integration will be when these taxonomic databases include references to the nomenclatural databases using LSIDs.

Chapter- 7

Phylogenetics

In biology, **phylogenetics** is the study of evolutionary relatedness among various groups of organisms (for example, species or populations), which is discovered through molecular sequencing data and morphological data matrices. The term *phylogenetics* is of Greek origin from the terms *phyle/phylon* (φυλή/φῦλον), meaning "tribe, race", and *genetikos* (γενετικός), meaning "relative to birth" from *genesis* (γένεσις, "birth"). Taxonomy, the classification, identification, and naming of organisms, has been richly informed by phylogenetics but remains methodologically and logically distinct.

The fields overlap however in the science of phylogenetic systematics – often called "cladism" or "cladistics" – where only phylogenetic trees are used to delimit taxa, which represent groups of lineage-connected individuals. In biological systematics as a whole, phylogenetic analyses have become essential in researching the evolutionary tree of life.

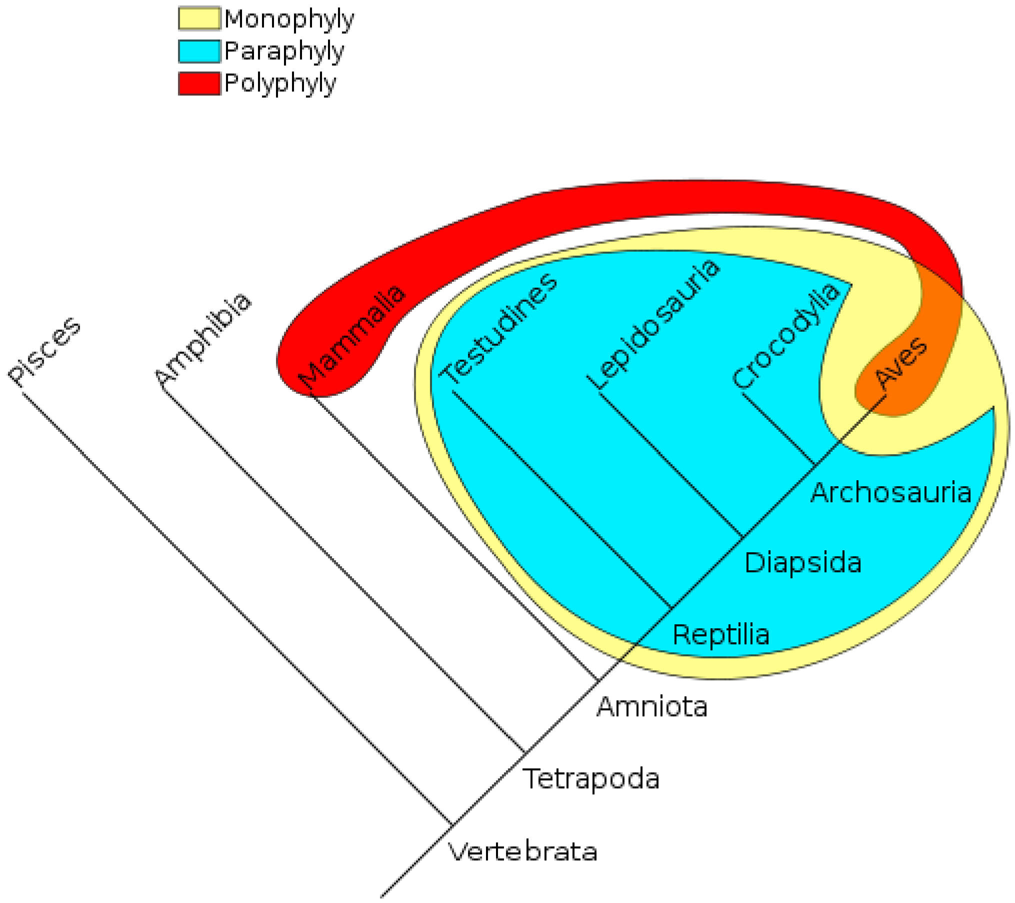
Construction of a phylogenetic tree

Evolution is regarded as a branching process, whereby populations are altered over time and may speciate into separate branches, hybridize together, or terminate by extinction. This may be visualized in a phylogenetic tree.

The problem posed by phylogenetics is that genetic data are only available for living taxa, and the fossil records (osteometric data) contains less data and more-ambiguous morphological characters. A phylogenetic tree represents a hypothesis of the order in which evolutionary events are assumed to have occurred.

Cladistics is the current method of choice to infer phylogenetic trees. The most commonly-used methods to infer phylogenies include parsimony, maximum likelihood, and MCMC-based Bayesian inference. Phenetics, popular in the mid-20th century but now largely obsolete, uses distance matrix-based methods to construct trees based on overall similarity, which is often assumed to approximate phylogenetic relationships. All methods depend upon an implicit or explicit mathematical model describing the evolution of characters observed in the species included, and are usually used for molecular phylogeny, wherein the characters are aligned nucleotide or amino acid sequences.

Grouping of organisms



Phylogenetic groups, or *taxa*, can be monophyletic, paraphyletic, or polyphyletic.

There are some terms that describe the nature of a grouping in such trees. For instance, all birds and reptiles are believed to have descended from a single common ancestor, so this taxonomic grouping (yellow in the diagram below) is called monophyletic. "Modern reptile" (cyan in the diagram) is a grouping that contains a common ancestor, but does not contain all descendants of that ancestor (birds are excluded).

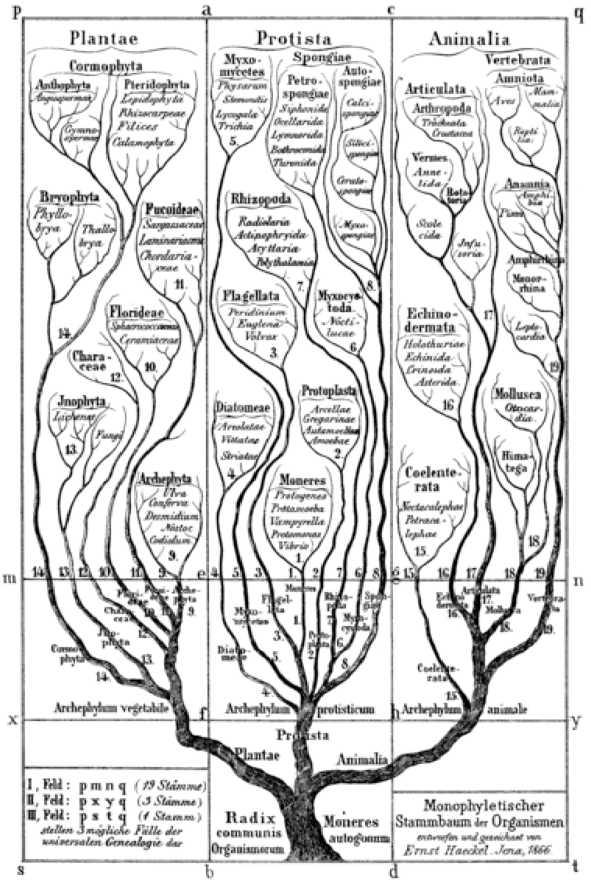
This is an example of a paraphyletic group. A grouping such as warm-blooded animals would include only mammals and birds (red/orange in the diagram) and is called polyphyletic because the members of this grouping do not include the most recent common ancestor.

Molecular phylogenetics

The evolutionary connections between organisms are represented graphically through phylogenetic trees. Due to the fact that evolution takes place over long periods of time that cannot be observed directly, biologists must reconstruct phylogenies by inferring the evolutionary relationships among present-day organisms. Fossils can aid with the reconstruction of phylogenies; however, fossil records are often too poor to be of good help. Therefore, biologists tend to be restricted with analysing present-day organisms to identify their evolutionary relationships. Phylogenetic relationships in the past were reconstructed by looking at phenotypes, often anatomical characteristics. Today, molecular data, which includes protein and DNA sequences, are used to construct phylogenetic trees.

The overall goal of National Science Foundation's Assembling the Tree of Life activity (AToL) is to resolve evolutionary relationships for large groups of organisms throughout the history of life, with the research often involving large teams working across institutions and disciplines. Investigators are typically supported for projects in data acquisition, analysis, algorithm development and dissemination in computational phylogenetics and phyloinformatics. For example, RedToL aims at reconstructing the Red Algal Tree of Life.

Ernst Haeckel's recapitulation theory



Genealogical tree suggested by Haeckel (1866)

During the late 19th century, Ernst Haeckel's recapitulation theory, or biogenetic law, was widely accepted. This theory was often expressed as "ontogeny recapitulates phylogeny", i.e. the development of an organism exactly mirrors the evolutionary development of the species. Haeckel's early version of this hypothesis [that the embryo mirrors *adult* evolutionary ancestors] has since been rejected, and the hypothesis amended as the embryo's development mirroring *embryos* of its evolutionary ancestors. He was accused by five professors of falsifying his images of embryos. Most modern biologists recognize numerous connections between ontogeny and phylogeny, explain them using evolutionary theory, or view them as supporting evidence for that theory. Donald I. Williamson suggested that larvae and embryos represented adults in other taxa that have been transferred by hybridization (the larval transfer theory). However, Williamson's views do not represent mainstream thought in molecular biology, and there is a significant body of evidence against the larval transfer theory.

Gene transfer

In general, organisms can inherit genes in two ways: vertical gene transfer and horizontal gene transfer. Vertical gene transfer is the passage of genes from parent to offspring, and horizontal gene transfer or lateral gene transfer occurs when genes jump between unrelated organisms, a common phenomenon in prokaryotes.

Horizontal gene transfer has complicated the determination of phylogenies of organisms, and inconsistencies in phylogeny have been reported among specific groups of organisms depending on the genes used to construct evolutionary trees.

Carl Woese came up with the three-domain theory of life (eubacteria, archaea and eukaryotes) based on his discovery that the genes encoding ribosomal RNA are ancient and distributed over all lineages of life with little or no horizontal gene transfer. Therefore, rRNAs are commonly recommended as molecular clocks for reconstructing phylogenies.

This has been particularly useful for the phylogeny of microorganisms, to which the species concept does not apply and which are too morphologically simple to be classified based on phenotypic traits.

Taxon sampling and phylogenetic signal

Owing to the development of advanced sequencing techniques in molecular biology, it has become feasible to gather large amounts of data (DNA or amino acid sequences) to infer phylogenetic hypotheses. For example, it is not rare to find studies with character matrices based on whole mitochondrial genomes (~16,000 nucleotides, in many animals). However, it has been proposed that it is more important to increase the number of taxa in the matrix than to increase the number of characters, because the more taxa the more robust is the resulting phylogenetic tree.

This may be partly due to the breaking up of long branches. It has been argued that this is an important reason to incorporate data from fossils into phylogenies where possible. Of course, phylogenetic data that include fossil taxa are generally based on morphology, rather than DNA data. Using simulations, Derrick Zwickl and David Hillis found that increasing taxon sampling in phylogenetic inference has a positive effect on the accuracy of phylogenetic analyses.

Another important factor that affects the accuracy of tree reconstruction is whether the data analyzed actually contain a useful phylogenetic signal, a term that is used generally to denote whether related organisms tend to resemble each other with respect to their genetic material or phenotypic traits. Ultimately, however, there is no way to measure whether a particular phylogenetic hypothesis is accurate or not, unless the "true" relationships among the taxa being examined are already known. The best result an empirical systematist can hope to attain is a tree with branches well-supported by the available evidence.

Importance of missing data

In general, the more data that is available when constructing a tree, the more accurate and reliable the resulting tree will be. Missing data is no less detrimental than simply having less data, although its impact is greatest when most of the missing data is in a small number of taxa. The fewer characters that have missing data, the better; concentrating the missing data across a small number of character states produces a more robust tree.

Role of fossils

Because many morphological characters involve embryological or soft-tissue characters that cannot be fossilized, and the interpretation of fossils is more ambiguous than living taxa, it is sometimes difficult to incorporate fossil data into phylogenies. However, despite these limitations, the inclusion of fossils is invaluable, as they can provide information in sparse areas of trees, breaking up long branches and constraining intermediate character states; thus, fossil taxa contribute as much to tree resolution as modern taxa.

Molecular phylogenies can reveal rates of diversification, but in order to track rates of origination, extinction and patterns in diversification, fossil data must be incorporated. Molecular techniques assume a constant rate of diversification, which is rarely likely to be true; in some (but by no means all) cases, the assumptions inherent in interpreting the fossil record (e.g. a complete and unbiased record) are closer to being true than the assumption of a constant rate, making fossil insights more accurate than molecular reconstructions.

Homoplasy weighting

Certain characters are more likely to be evolved convergently than others; logically, such characters should be given less weight in the reconstruction of a tree. Unfortunately the only objective way to determine convergence is by the construction of a tree – a somewhat circular method. Even so, weighting homoplasious characters does indeed lead to better-supported trees. Further refinement can be brought by weighting changes in one direction higher than changes in another; for instance, the presence of thoracic wings almost guarantees placement among the pterygote insects, although because wings are often lost secondarily, their absence does not exclude a taxon from the group.

Computational Phylogenetics

Computational phylogenetics is the application of computational algorithms, methods and programs to phylogenetic analyses. The goal is to assemble a phylogenetic tree representing a hypothesis about the evolutionary ancestry of a set of genes, species, or other taxa. For example, these techniques have been used to explore the family tree of hominid species and the relationships between specific genes shared by many types of organisms. Traditional phylogenetics relies on morphological data obtained by measuring and quantifying the phenotypic properties of representative organisms, while the more recent field of molecular phylogenetics uses nucleotide sequences encoding genes or amino acid sequences encoding proteins as the basis for classification. Many forms of molecular phylogenetics are closely related to and make extensive use of sequence alignment in constructing and refining phylogenetic trees, which are used to classify the evolutionary relationships between homologous genes represented in the genomes of divergent species. The phylogenetic trees constructed by computational methods are unlikely to perfectly reproduce the evolutionary tree that represents the historical relationships between the species being analyzed. The historical species tree may also differ from the historical tree of an individual homologous gene shared by those species.

Producing a phylogenetic tree requires a measure of homology among the characteristics shared by the taxa being compared. In morphological studies, this requires explicit decisions about which physical characteristics to measure and how to use them to encode distinct states corresponding to the input taxa. In molecular studies, a primary problem is in producing a multiple sequence alignment (MSA) between the genes or amino acid sequences of interest. Progressive sequence alignment methods produce a phylogenetic tree by necessity because they incorporate new sequences into the calculated alignment in order of genetic distance.

Types of phylogenetic trees

Phylogenetic trees generated by computational phylogenetics can be either *rooted* or *unrooted* depending on the input data and the algorithm used. A rooted tree is a directed graph that explicitly identifies a most recent common ancestor (MRCA), usually an imputed sequence that is not represented in the input. Genetic distance measures can be used to plot a tree with the input sequences as leaf nodes and their distances from the root proportional to their genetic distance from the hypothesized MRCA. Identification of a root usually requires the inclusion in the input data of at least one "outgroup" known to be only distantly related to the sequences of interest.

By contrast, unrooted trees plot the distances and relationships between input sequences without making assumptions regarding their descent. An unrooted tree can always be produced from a rooted tree, but a root cannot usually be placed on an unrooted tree without additional data on divergence rates, such as the assumption of the molecular clock hypothesis.

The set of all possible phylogenetic trees for a given group of input sequences can be conceptualized as a discretely defined multidimensional "tree space" through which search paths can be traced by optimization algorithms. Although counting the total number of trees for a nontrivial number of input sequences can be complicated by variations in the definition of a tree topology, it is always true that there are more rooted than unrooted trees for a given number of inputs and choice of parameters.

Coding characters and defining homology

Morphological analysis

The basic problem in morphological phylogenetics is the assembly of a matrix representing a mapping from each of the taxa being compared to representative measurements for each of the phenotypic characteristics being used as a classifier. The types of phenotypic data used to construct this matrix depend on the taxa being compared; for individual species, they may involve measurements of average body size, lengths or sizes of particular bones or other physical features, or even behavioral manifestations. Of course, since not every possible phenotypic characteristic could be measured and encoded for analysis, the selection of which features to measure is a major inherent obstacle to the method. The decision of which traits to use as a basis for the matrix necessarily represents a hypothesis about which traits of a species or higher taxon are evolutionarily relevant. Morphological studies can be confounded by examples of convergent evolution of phenotypes. A major challenge in constructing useful classes is the high likelihood of inter-taxon overlap in the distribution of the phenotype's variation. The inclusion of extinct taxa in morphological analysis is often difficult due to absence of or incomplete fossil records, but has been shown to have a significant effect on the trees produced; in one study only the inclusion of extinct species of apes produced a morphologically derived tree that was consistent with that produced from molecular data.

Some phenotypic classifications, particularly those used when analyzing very diverse groups of taxa, are discrete and unambiguous; classifying organisms as possessing or lacking a tail, for example, is straightforward in the majority of cases, as is counting features such as eyes or vertebrae. However, the most appropriate representation of continuously varying phenotypic measurements is a controversial problem without a general solution. A common method is simply to sort the measurements of interest into two or more classes, rendering continuous observed variation as discretely classifiable (e.g., all examples with humerus bones longer than a given cutoff are scored as members of one state, and all members whose humerus bones are shorter than the cutoff are scored as members of a second state). This results in an easily manipulated data set but has been criticized for poor reporting of the basis for the class definitions and for sacrificing information compared to methods that use a continuous weighted distribution of measurements.

Because morphological data is extremely labor-intensive to collect, whether from literature sources or from field observations, reuse of previously compiled data matrices is not uncommon, although this may propagate flaws in the original matrix into multiple derivative analyses.

Molecular analysis

The problem of character coding is very different in molecular analyses, as the characters in biological sequence data are immediate and discretely defined - distinct nucleotides in DNA or RNA sequences and distinct amino acids in protein sequences. However, defining homology can be challenging due to the inherent difficulties of multiple sequence alignment. For a given gapped MSA, several rooted phylogenetic trees can be constructed that vary in their interpretations of which changes are "mutations" versus ancestral characters, and which events are insertion mutations or deletion mutations. For example, given only a pairwise alignment with a gap region, it is impossible to determine whether one sequence bears an insertion mutation or the other carries a deletion. The problem is magnified in MSAs with unaligned and nonoverlapping gaps. In practice, sizable regions of a calculated alignment may be discounted in phylogenetic tree construction to avoid integrating noisy data into the tree calculation.

Distance-matrix methods

Distance-matrix methods of phylogenetic analysis explicitly rely on a measure of "genetic distance" between the sequences being classified, and therefore they require an MSA as an input. Distance is often defined as the fraction of mismatches at aligned positions, with gaps either ignored or counted as mismatches. Distance methods attempt to construct an all-to-all matrix from the sequence query set describing the distance between each sequence pair. From this is constructed a phylogenetic tree that places closely related sequences under the same interior node and whose branch lengths closely reproduce the observed distances between sequences. Distance-matrix methods may produce either rooted or unrooted trees, depending on the algorithm used to calculate them. They are frequently used as the basis for progressive and iterative types of multiple sequence alignments. The main disadvantage of distance-matrix methods is their inability to efficiently use information about local high-variation regions that appear across multiple subtrees.

Neighbor-joining

Neighbor-joining methods apply general data clustering techniques to sequence analysis using genetic distance as a clustering metric. The simple neighbor-joining method produces unrooted trees, but it does not assume a constant rate of evolution (i.e., a molecular clock) across lineages. Its relative, UPGMA (Unweighted Pair Group Method with Arithmetic mean) produces rooted trees and requires a constant-rate assumption - that is, it assumes an ultrametric tree in which the distances from the root to every branch tip are equal.

Fitch-Margoliash method

The Fitch-Margoliash method uses a weighted least squares method for clustering based on genetic distance. Closely related sequences are given more weight in the tree construction process to correct for the increased inaccuracy in measuring distances between distantly related sequences. The distances used as input to the algorithm must be normalized to prevent large artifacts in computing relationships between closely related and distantly related groups. The distances calculated by this method must be linear; the linearity criterion for distances requires that the expected values of the branch lengths for two individual branches must equal the expected value of the sum of the two branch distances - a property that applies to biological sequences only when they have been corrected for the possibility of back mutations at individual sites. This correction is done through the use of a substitution matrix such as that derived from the Jukes-Cantor model of DNA evolution. The distance correction is only necessary in practice when the evolution rates differ among branches.

The least-squares criterion applied to these distances is more accurate but less efficient than the neighbor-joining methods. An additional improvement that corrects for correlations between distances that arise from many closely related sequences in the data set can also be applied at increased computational cost. Finding the optimal least-squares tree with any correction factor is NP-complete, so heuristic search methods like those used in maximum-parsimony analysis are applied to the search through tree space.

Using outgroups

Independent information about the relationship between sequences or groups can be used to help reduce the tree search space and root unrooted trees. Standard usage of distance-matrix methods involves the inclusion of at least one outgroup sequence known to be only distantly related to the sequences of interest in the query set. This usage can be seen as a type of experimental control. If the outgroup has been appropriately chosen, it will have a much greater genetic distance and thus a longer branch length than any other sequence, and it will appear near the root of a rooted tree. Choosing an appropriate outgroup requires the selection of a sequence that is moderately related to the sequences of interest; too close a relationship defeats the purpose of the outgroup and too distant adds noise to the analysis. Care should also be taken to avoid situations in which the species from which the sequences were taken are distantly related, but the gene encoded by the sequences is highly conserved across lineages. Horizontal gene transfer, especially between otherwise divergent bacteria, can also confound outgroup usage.

Maximum parsimony

Maximum parsimony (MP) is a method of identifying the potential phylogenetic tree that requires the smallest total number of evolutionary events to explain the observed sequence data. Some ways of scoring trees also include a "cost" associated with particular types of evolutionary events and attempt to locate the tree with the smallest total cost. This is a useful approach in cases where not every possible type of event is equally likely - for example, when particular nucleotides or amino acids are known to be more mutable than others.

The most naive way of identifying the most parsimonious tree is simple enumeration - considering each possible tree in succession and searching for the tree with the smallest score. However, this is only possible for a relatively small number of sequences or species because the problem of identifying the most parsimonious tree is known to be NP-hard; consequently a number of heuristic search methods for optimization have been developed to locate a highly parsimonious tree, if not the most optimal in the set. Most such methods involve a steepest descent-style minimization mechanism operating on a tree rearrangement criterion.

Branch and bound

The branch and bound algorithm is a general method used to increase the efficiency of searches for near-optimal solutions of NP-hard problems first applied to phylogenetics in the early 1980s. Branch and bound is particularly well suited to phylogenetic tree construction because it inherently requires dividing a problem into a tree structure as it subdivides the problem space into smaller regions. As its name implies, it requires as input both a branching rule (in the case of phylogenetics, the addition of the next species or sequence to the tree) and a bound (a rule that excludes certain regions of the search space from consideration, thereby assuming that the optimal solution cannot occupy that region). Identifying a good bound is the most challenging aspect of the algorithm's application to phylogenetics. A simple way of defining the bound is a maximum number of assumed evolutionary changes allowed per tree. A set of criteria known as Zharkikh's rules severely limit the search space by defining characteristics shared by all candidate "most parsimonious" trees. The two most basic rules require the elimination of all but one redundant sequence (for cases where multiple observations have produced identical data) and the elimination of character sites at which two or more states do not occur in at least two species. Under ideal conditions these rules and their associated algorithm would completely define a tree.

Sankoff-Morel-Cedergren algorithm

The Sankoff-Morel-Cedergren algorithm was among the first published methods to simultaneously produce an MSA and a phylogenetic tree for nucleotide sequences. The method uses a maximum parsimony calculation in conjunction with a scoring function that penalizes gaps and mismatches, thereby favoring the tree that introduces a minimal number of such events. The imputed sequences at the interior nodes of the tree are scored and summed over all the nodes in each possible tree. The lowest-scoring tree sum provides both an optimal tree and an optimal MSA given the scoring function. Because the method is highly computationally intensive, an approximate method in which initial guesses for the interior alignments are refined one node at a time. Both the full and the approximate version are in practice calculated by dynamic programming.

MALIGN and POY

More recent phylogenetic tree/MSA methods use heuristics to isolate high-scoring, but not necessarily optimal, trees. The MALIGN method uses a maximum-parsimony technique to compute a multiple alignment by maximizing a cladogram score, and its companion POY uses an iterative method that couples the optimization of the phylogenetic tree with improvements in the corresponding MSA. However, the use of these methods in constructing evolutionary hypotheses has been criticized as biased due to the deliberate construction of trees reflecting minimal evolutionary events.

Maximum likelihood

The maximum likelihood method uses standard statistical techniques for inferring probability distributions to assign probabilities to particular possible phylogenetic trees. The method requires a substitution model to assess the probability of particular mutations; roughly, a tree that requires more mutations at interior nodes to explain the observed phylogeny will be assessed as having a lower probability. This is broadly similar to the maximum-parsimony method, but maximum likelihood allows additional statistical flexibility by permitting varying rates of evolution across both lineages and sites. In fact, the method requires that evolution at different sites and along different lineages must be statistically independent. Maximum likelihood is thus well suited to the analysis of distantly related sequences, but because it formally requires search of all possible combinations of tree topology and branch length, it is computationally expensive to perform on more than a few sequences.

The "pruning" algorithm, a variant of dynamic programming, is often used to reduce the search space by efficiently calculating the likelihood of subtrees. The method calculates the likelihood for each site in a "linear" manner, starting at a node whose only descendants are leaves (that is, the tips of the tree) and working backwards toward the "bottom" node in nested sets. However, the trees produced by the method are only rooted if the substitution model is irreversible, which is not generally true of biological systems. The search for the maximum-likelihood tree also includes a branch length optimization component that is difficult to improve upon algorithmically; general global optimization tools such as the Newton-Raphson method are often used. Searching tree topologies defined by likelihood has not been shown to be NP-complete, but remains extremely challenging because branch-and-bound search is not yet effective for trees represented in this way.

Bayesian inference

Bayesian inference can be used to produce phylogenetic trees in a manner closely related to the maximum likelihood methods. Bayesian methods assume a prior probability distribution of the possible trees, which may simply be the probability of any one tree among all the possible trees that could be generated from the data, or may be a more sophisticated estimate derived from the assumption that divergence events such as speciation occur as stochastic processes. The choice of prior distribution is a point of contention among users of Bayesian-inference phylogenetics methods.

Implementations of Bayesian methods generally use Markov chain Monte Carlo sampling algorithms, although the choice of move set varies; selections used in Bayesian phylogenetics include circularly permuting leaf nodes of a proposed tree at each step and swapping descendant subtrees of a random internal node between two related trees. The use of Bayesian methods in phylogenetics has been controversial, largely due to incomplete specification of the choice of move set, acceptance criterion, and prior distribution in published work.

Model selection

Molecular phylogenetics methods rely on a defined substitution model that encodes a hypothesis about the relative rates of mutation at various sites along the gene or amino acid sequences being studied. At their simplest, substitution models aim to correct for differences in the rates of transitions and transversions in nucleotide sequences. The use of substitution models is necessitated by the fact that the genetic distance between two sequences increases linearly only for a short time after the two sequences diverge from each other (alternatively, the distance is linear only shortly before coalescence). The longer the amount of time after divergence, the more likely it becomes that two mutations occur at the same nucleotide site. Simple genetic distance calculations will thus undercount the number of mutation events that have occurred in evolutionary history. The extent of this undercount increases with increasing time since divergence, which can lead to the phenomenon of long branch attraction, or the misassignment of two distantly related but convergently evolving sequences as closely related. The maximum parsimony method is particularly susceptible to this problem due to its explicit search for a tree representing a minimum number of distinct evolutionary events.

Types of models

All substitution models assign a set of weights to each possible change of state represented in the sequence. The most common model types are implicitly reversible because they assign the same weight to, for example, a G>C nucleotide mutation as to a C>G mutation. The simplest possible model, the Jukes-Cantor model, assigns an equal probability to every possible change of state for a given nucleotide base. The rate of change between any two distinct nucleotides will be one-third of the overall substitution rate. More advanced models distinguish between transitions and transversions. The most general possible time-reversible model, called the GTR model, has six mutation rate parameters. An even more generalized model known as the general 12-parameter model breaks time-reversibility, at the cost of much additional complexity in calculating genetic distances that are consistent among multiple lineages. One possible variation on this theme adjusts the rates so that overall GC content - an important measure of DNA double helix stability - varies over time.

Models may also allow for the variation of rates with positions in the input sequence. The most obvious example of such variation follows from the arrangement of nucleotides in protein-coding genes into three-base codons. If the location of the open reading frame (ORF) is known, rates of mutation can be adjusted for position of a given site within a codon, since it is known that wobble base pairing can allow for higher mutation rates in the third nucleotide of a given codon without affecting the codon's meaning in the genetic code. A less hypothesis-driven example that does not rely on ORF identification simply assigns to each site a rate randomly drawn from a predetermined distribution, often the gamma distribution or log-normal distribution. Finally, a more conservative estimate of rate variations known as the covarion method allows autocorrelated variations in rates, so that the mutation rate of a given site is correlated across sites and lineages.

Choosing the best model

The selection of an appropriate model is critical for the production of good phylogenetic analyses, both because underparameterized or overly restrictive models may produce aberrant behavior when their underlying assumptions are violated, and because overly complex or overparameterized models are computationally expensive and the parameters may be overfit. The most common method of model selection is the likelihood ratio test (LRT), which produces a likelihood estimate that can be interpreted as a measure of "goodness of fit" between the model and the input data. However, care must be taken in using these results, since a more complex model with more parameters will always have a higher likelihood than a simplified version of the same model, which can lead to the naive selection of models that are overly complex. For this reason model selection computer programs will choose the simplest model that is not significantly worse than more complex substitution models. A significant disadvantage of the LRT is the necessity of making a series of pairwise comparisons between models; it has been shown that the order in which the models are compared has a major effect on the one that is eventually selected.

An alternative model selection method is the Akaike information criterion (AIC), formally an estimate of the Kullback-Leibler divergence between the true model and the model being tested. It can be interpreted as a likelihood estimate with a correction factor to penalize overparameterized models. The AIC is calculated on an individual model rather than a pair, so it is independent of the order in which models are assessed. A related alternative, the Bayesian information criterion (BIC), has a similar basic interpretation but penalizes complex models more heavily.