**Course: Advanced Bioinformatics**

**Module title: Phylogeny Analysis**

**Module no. : 50**

Cladistics is an approach to biological classification in which organisms are grouped together based on whether or not they have one or more shared unique characteristics that come from the group's last common ancestor and are not present in more distant ancestors. Therefore, members of the same group are thought to share a common history and are considered to be more closely related.

The original methods used in cladistics analysis and the school of taxonomy derived from the work of the German entomologist Willi Hennig, who referred to it as phylogenetic systematics (also the title of his 1966 book); the terms "cladistics" and "clade" were popularized by other researchers. Cladistics in the original sense refers to a particular set of methods used in phylogenetic analysis, although it is now sometimes used to refer to the whole field.

**Methodology:**

The cladistic method interprets each character state transformation implied by the distribution of shared character states among taxa (or other terminals) as a potential piece of evidence for grouping. The outcome of a cladistic analysis is a [cladogram](https://en.wikipedia.org/wiki/Cladogram) – a [tree](https://en.wikipedia.org/wiki/Tree_(graph_theory))-shaped diagram ([dendrogram](https://en.wikipedia.org/wiki/Dendrogram" \o "Dendrogram))[[17]](https://en.wikipedia.org/wiki/Cladistics" \l "cite_note-17) that is interpreted to represent the best hypothesis of phylogenetic relationships. Although traditionally such cladograms were generated largely on the basis of morphological characters and originally calculated by hand, [genetic sequencing](https://en.wikipedia.org/wiki/DNA_sequencing) data and [computational phylogenetics](https://en.wikipedia.org/wiki/Computational_phylogenetics) are now commonly used in phylogenetic analyses, and the[parsimony](https://en.wikipedia.org/wiki/Parsimony#Biology) criterion has been abandoned by many phylogeneticists in favor of more "sophisticated" but less parsimonious evolutionary models of character state transformation. Cladists contend that these models are unjustified.

Every cladogram is based on a particular dataset analyzed with a particular method. Datasets are tables consisting of [molecular](https://en.wikipedia.org/wiki/Molecular_phylogenetics), morphological, [ethological](https://en.wikipedia.org/wiki/Ethology)[[18]](https://en.wikipedia.org/wiki/Cladistics#cite_note-18) and/or other characters and a list of [operational taxonomic units](https://en.wikipedia.org/wiki/Operational_taxonomic_unit) (OTUs), which may be genes, individuals, populations, species, or larger taxa that are presumed to be monophyletic and therefore to form, all together, one large clade; phylogenetic analysis infers the branching pattern within that clade. Different datasets and different methods, not to mention violations of the mentioned assumptions, often result in different cladograms. Only scientific investigation can show which is more likely to be correct.

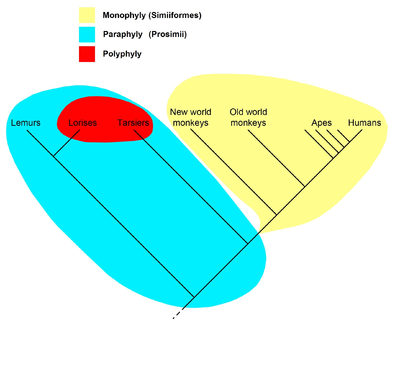
Until recently, for example, cladograms like the following have generally been accepted as accurate representations of the ancestral relations among turtles, lizards, crocodilians, and birds

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ▼ | |  |  | | --- | --- | | [Testudines](https://en.wikipedia.org/wiki/Turtle) | [turtles](https://en.wikipedia.org/wiki/Turtle) | |  | | [Diapsida](https://en.wikipedia.org/wiki/Diapsid)   ♦ | |  |  | | --- | --- | | [Lepidosauria](https://en.wikipedia.org/wiki/Lepidosauria) | [lizards](https://en.wikipedia.org/wiki/Lizard) | |  | | [Archosauria](https://en.wikipedia.org/wiki/Archosaur) | |  |  | | --- | --- | | [Crocodylomorpha](https://en.wikipedia.org/wiki/Crocodylomorpha) | [crocodilians](https://en.wikipedia.org/wiki/Crocodilia) | |  | | [Dinosauria](https://en.wikipedia.org/wiki/Dinosaur) | [birds](https://en.wikipedia.org/wiki/Bird) | |  | | |  | | |  | |
|  |

If this phylogenetic hypothesis is correct, then the last common ancestor of turtles and birds, at the branch near the ▼ lived earlier than the last common ancestor of lizards and birds, near the ♦. Most [molecular](https://en.wikipedia.org/wiki/Molecular_phylogenetics" \o "Molecular phylogenetics)evidence, however, produces cladograms more like this:[[20]](https://en.wikipedia.org/wiki/Cladistics#cite_note-20)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| [Diapsida](https://en.wikipedia.org/wiki/Diapsid)   ♦ | |  |  | | --- | --- | | [Lepidosauria](https://en.wikipedia.org/wiki/Lepidosauria) | [lizards](https://en.wikipedia.org/wiki/Lizard) | |  | | Lepidosauria▼ | |  |  | | --- | --- | | [Testudines](https://en.wikipedia.org/wiki/Turtle) | [turtles](https://en.wikipedia.org/wiki/Turtle) | |  | | [Archosauria](https://en.wikipedia.org/wiki/Archosaur) | |  |  | | --- | --- | | [Crocodylomorpha](https://en.wikipedia.org/wiki/Crocodylomorpha) | [crocodilians](https://en.wikipedia.org/wiki/Crocodilia) | |  | | [Dinosauria](https://en.wikipedia.org/wiki/Dinosaur) | [birds](https://en.wikipedia.org/wiki/Bird) | |  | | |  | | |  | |
|  |

If this is accurate, then the last common ancestor of turtles and birds lived later than the last common ancestor of lizards and birds. Since the cladograms provide competing accounts of real events, at most one of them is correct.

[](https://en.wikipedia.org/wiki/File:Monophyly,_paraphyly,_polyphyly.png)

Cladogram of the [primates](https://en.wikipedia.org/wiki/Primate), showing a [monophyletic](https://en.wikipedia.org/wiki/Monophyly) taxon (a [clade](https://en.wikipedia.org/wiki/Clade): the simians or Anthropoidea, in yellow), a [paraphyletic](https://en.wikipedia.org/wiki/Paraphyly) taxon (the prosimians, in blue, including the red patch), and a [polyphyletic](https://en.wikipedia.org/wiki/Polyphyly) taxon (the nocturnal primates – the [lorises](https://en.wikipedia.org/wiki/Loris" \o "Loris) and the [tarsiers](https://en.wikipedia.org/wiki/Tarsier) – in red)

The cladogram to the right represents the current universally accepted hypothesis that all [primates](https://en.wikipedia.org/wiki/Primate), including [strepsirrhines](https://en.wikipedia.org/wiki/Strepsirrhini" \o "Strepsirrhini) like the [lemurs](https://en.wikipedia.org/wiki/Lemur) and [lorises](https://en.wikipedia.org/wiki/Loris" \o "Loris), had a common ancestor all of whose descendants were primates, and so form a clade; the name Primates is therefore recognized for this clade. Within the primates, all anthropoids (monkeys, apes and humans) are hypothesized to have had a common ancestor all of whose descendants were anthropoids, so they form the clade called Anthropoidea. The "prosimians", on the other hand, form a paraphyletic taxon. The name Prosimii is not used in [phylogenetic nomenclature](https://en.wikipedia.org/wiki/Phylogenetic_nomenclature), which names only clades; the "prosimians" are instead divided between the clades [Strepsirhini](https://en.wikipedia.org/wiki/Strepsirrhini" \o "Strepsirrhini) and [Haplorhini](https://en.wikipedia.org/wiki/Haplorhini" \o "Haplorhini), where the latter contains Tarsiiformes and Anthropoidea.

**Terminology for taxa**

Mono-, para- and polyphyletic taxa can be understood based on the shape of the tree, as well as based on their character states.

|  |  |  |
| --- | --- | --- |
| **Term** | **Node-based definition** | **Character-based definition** |
| [Monophyly](https://en.wikipedia.org/wiki/Monophyly) | A [clade](https://en.wikipedia.org/wiki/Clade), a monophyletic taxon, is a taxon that includes all descendants of an inferred ancestor. | A clade is characterized by one or more **apomorphies**: derived character states present in the first member of the taxon, inherited by its descendants (unless secondarily lost), and not inherited by any other taxa. |
| [Paraphyly](https://en.wikipedia.org/wiki/Paraphyly) | A paraphyletic assemblage is one that is constructed by taking a clade and removing one or more smaller clades.[[28]](https://en.wikipedia.org/wiki/Cladistics#cite_note-28) (Removing one clade produces a singly paraphyletic assemblage, removing two produces a doubly paraphylectic assemblage, and so on.)[[29]](https://en.wikipedia.org/wiki/Cladistics#cite_note-29) | A paraphyletic assemblage is characterized by one or more **plesiomorphies**: character states inherited from ancestors but not present in all of their descendants. As a consequence, a paraphyletic assemblage is truncated, in that it excludes one or more clades from an otherwise monophyletic taxon. An alternative name is [*evolutionary grade*](https://en.wikipedia.org/wiki/Evolutionary_grade), referring to an ancestral character state within the group. While paraphyletic assemblages are popular among paleontologists and evolutionary taxonomists, cladists do not recognize paraphyletic assemblages as having any formal information content – they are merely parts of clades. |
| [Polyphyly](https://en.wikipedia.org/wiki/Polyphyly) | A polyphyletic assemblage is one which is neither monophyletic nor paraphyletic. | A polyphyletic assemblage is characterized by one or more **[homoplasies](https://en.wikipedia.org/wiki/Homoplasy" \o "Homoplasy)**: character states which have converged or reverted so as to be the same but which have not been inherited from a common ancestor. No systematist recognizes polyphyletic assemblages as taxonomically meaningful entities, although ecologists sometimes consider them meaningful labels for functional participants in ecological communities (e. g., primary producers, detritivores, etc.). |

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 (which may happen with bacteria or viruses under laboratory conditions). The best result an empirical phylogeneticist can hope to attain is a tree with branches that are well supported by the available evidence.

**Homoplasy**

Certain characters are more likely to evolve convergent than others; logically, such characters should be given less weight in the reconstruction of a tree. Weights in the form of a model of evolution can be inferred from sets of molecular data, so that maximum likelihood or Bayesian methods can be used to analyze them. For molecular sequences, this problem is exacerbated when the taxa under study have diverged substantially. As time since the divergence of two taxa increase, so does the probability of multiple substitutions on the same site, or back mutations, all of which result in homoplasies. For morphological data, unfortunately, the only objective way to determine convergence is by the construction of a tree – a somewhat circular method. Even so, weighting homoplasious characters [how?] 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 because, although wings are often lost secondarily, there is no evidence that they have been gained more than once.

**Horizontal 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 (also called lateral) gene transfer occurs when genes jump between unrelated organisms, a common phenomenon especially in prokaryotes; a good example of this is the acquired antibiotic resistance as a result of gene exchange between various bacteria leading to multi-drug-resistant bacterial species. There have also been well-documented cases of horizontal gene transfer between eukaryotes.

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. The only way to determine which genes have been acquired vertically and which horizontally is to parsimoniously assume that the largest set of genes that have been inherited together have been inherited vertically; this requires analyzing a large number of genes.

**Hybrids, speciation and introgressions**

The basic assumption underlying the mathematical model of cladistics is a situation where species split neatly in bifurcating fashion. While such an assumption may hold on a larger scale, speciation is often much less orderly. Research since the cladistic method was introduced has shown that hybrid speciation, once thought rare, is in fact quite common, particularly in plants. Also paraphyletic speciation is common, making the assumption of a bifurcating pattern unsuitable, leading to phylogenetic networks rather than trees. Introgression can also move genes between otherwise distinct species and sometimes even genera, complicating phylogenetic analysis based on genes. This phenomenon can contribute to "incomplete line sorting" and is thought to be a common phenomenon across a number of groups. In species level analysis this can be dealt with by larger sampling or better whole genome analysis. Often the problem is avoided by restricting the analysis to fewer, not closely related specimen.

Taxon sampling

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, simulations have shown 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 there are, the more accurate and more robust is the resulting phylogenetic tree. This may be partly due to the breaking up of long branches.

Phylogenetic signal

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 a character evolves slowly enough to have the same state in closely related taxa as opposed to varying randomly. Tests for phylogenetic signal exist.