Character Weighting and Coding

Polarity

Polarity is the designation of relative ancestry to the character states of a morphocline. A change in character state represents a heritable evolutionary modification from a pre-existing structure or feature (termed **plesiomorphic**, ancestral, or primitive) to a new structure or feature (termed **apomorphic**, derived, or advanced).

The designation of polarity is often one of the uncertain aspects of phylogenetic analysis. The methods used for polarity designation are:

- comparison may be made within the concerned group shown in green (**in-group comparison**) or
- with relatives outside the group shown in pink (**outgroup comparison**).

Character Coding and Weighting

The **coding** of character states is done by assigning non-negative integer values.

Binary or Two-state coding

Binary characters are conveniently assigned 0 and 1 for two states. If it is possible to distinguish, plesiomorphic state is assigned 0 and apomorphic state 1 code

Multistate Characters

Multiatate characters can be coded into number of states using non-negative integer numbers (0,1,2,3,4…) corresponding to transformation series (ordered or unordered).

As part of a phylogenetic analysis, the characters may be assigned weights. **Character weighting** is the assignment of greater or lesser taxonomic importance to certain characters over other characters in determining phylogenetic relationships.

More commonly the numerical taxonomy approach to give **equal weighting** to all the characters is adopted in the preliminary analysis, then characters which show the least homoplasy (similarity that is *not* the result of homology is termed **homoplasy or analogy**) are identified and given more weightage in the subsequent analysis, this process is known as **successive weighting**. Thus characters that exhibit reversals

(shown by red arrow) or parallelisms (shown by green arrows) are recognized and given less weight over those that do not.

Character X Taxon Matrix

The coded data may be organized in the form of a matrix or table with ***t*** number of rows (OTUs) and ***n*** number of columns (character-states). The character x taxon matrix supplies the data for constructing a phylogenetic tree or cladogram.

Once the data have been codified and entered in the form of a matrix, the next step is to calculate the degree of similarity between every pair of OTUs. A number of formulae have been proposed by various authors to calculate **similarity** or **dissimilarity** (**taxonomic distance**) between the OTUs. :

- Simple matching coefficient
- Jaccard Coefficient of association
- Yule coefficient

Once the similarity or distance between every pair of taxa has been calculated, the data are presented in a second matrix with **t** x **t** dimensions where both rows and columns represent taxa. It must be noted that diagonal **t** value in the matrix represents self-comparison of taxa and thus 100% similarity. These values are redundant as such. The values in the triangle above this diagonal line would be similar to the triangle below.

Construction of Trees

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

Parsimony-based methods

Because there are generally many possible trees for any given data set, one of the major methods of reconstructing phylogenetic relationships is known as the **principle of parsimony** or **parsimony analysis**. **The principle of parsimony states that of the numerous possible cladograms for a given group of OTUs, the one (or more) exhibiting the fewest number of evolutionary steps is accepted as being the best estimate of**

phylogeny. (Note that there may be two or more cladograms that are equally most parsimonious.)

Distance methods

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

Cluster analysis

Data presented in OTUs x OTUs (**t** x **t**) matrix are too exhaustive to provide any meaningful picture and need to be further condensed to enable a comparison of units. Cluster analysis is one such method in which OTUs are arranged in the order of decreasing similarity. Computer programs are used to efficient cluster analysis and help in the construction of cluster diagrams or phenograms. The various clustering procedures are classified under two categories:

Agglomerative methods

Agglomerative methods start with **t** clusters equal to the number of OTUs. These are successively merged until a single cluster has finally been formed. The most commonly used clustering method in biology is **Sequential Agglomerative Hierarchic Non-overlapping clustering method** (SAHN). The method is useful for achieving hierarchical classifications.

Divisive methods

Divisive methods as opposed to agglomerative methods, start with all **t,** OTUs as a single set, subdividing this into one or more subsets; this is continued until further subdivision is not necessary. The commonly used divisive method is **association**. The method has been mostly used in ecological data employing two state characters. It builds a dendrogram from the top downwards as opposed to cluster analysis, which builds a diagram from the bottom up.

The results of cluster analysis are commonly presented as dendrograms known as **phenograms**.

Phenograms and Cladograms

Phenetics

Phenetics is the sudy of relationships among a group of organisms on the basis of degree of similarity between them. A tree like diagram that represents phenetic relationships is called a phenogram.

Cladistics

Phylogeny refers to the evolutionary history or pattern of descent of a group of organisms. Phylogenetic systematics, or cladistics, is that branch of systematics concerned with inferring phylogeny. A tree like diagram that represents ancestor-descendant relationships is called a **cladogram**.

Phenetics	Cladistics
Phenetics uses overall similarity – all characters are used.	Cladistic uses only phylogenetically informative characters.
Current state of characters is considered without assigning any polarity.	Primitive and derived characters states must be identified and assigned a polarity.
Taxa may be grouped together by shared *ancestral* features (known as symplesiomorphies) as well as by shared derived character states (synapomorphies).	Taxa may be grouped together by shared derived character states (synapomorphies) only.
An outgroup is not included in analysis.	An outgroup must be included in the analysis