March 12, 2018. **Classification I -- introduction to phylogenetic classifications; monophyly, information content**

Reading: *Tree Thinking*: pp 107-131

Classifications are required to provide a context for biodiversity in scientific studies. Thus given the importance of taxonomic classification, the criteria for these classifications should rest on the highest quality of data available, and if these data change the classification should be adjusted accordingly. We expect this of the models used in quantum physics, molecular biology, health sciences, or engineering, and we should expect nothing less for biodiversity estimates and comparisons.

"Classification" versus "taxonomy" versus "systematics" versus "nomenclature"?

The debate over classification has a long and checkered history (see Hull 1988; Stevens, 1994; Mishler 2013). A conceptual upheaval in the 1970's and 80's resulted in a true scientific revolution --Hennigian Phylogenetic Systematics. Many issues were at stake in that era, foremost of which was the nature of taxa. Are they just convenient groupings of organisms with similar features, or are they lineages, marked by homologies?

"Natural classification"

All three schools of systematics wanted to produce "natural" classifications:

- **Pheneticists** view natural groups as those taxa linked by the greatest similarity to each other.
- **Evolutionary systematists** view "natural" groups as defined by gaps between taxa in characters for which an evolutionary scenario can be argued.
- **Cladists** consider natural groups to be monophyletic, and thus "natural" classifications reflect the tree of life. Cladism is the approach to classification that defines taxa by uniquely shared common ancestry (monophyly), as evidenced by shared derived characters. It is also known as phylogenetic systematics.

As discussed in the Jan 22nd lecture, properties of a ideal classification system include: (1) information content (summarizing what is known about organisms), (2) predictivity (what is not yet known about organisms), and (3) function in theories (capturing entities involved in important natural processes). A general, if not completely universal consensus has been reached, that taxa are (or at least should be) phylogenetic (Hennig, 1966; Nelson, 1973; Farris, 1983; Sober, 1988). In phylogenetic classifications, the last mentioned criterion tracks causal relations (e.g., evolution from common ancestors) even in the absence of detailed knowledge of those causes.

A summary of the arguments for why formal taxonomic names should be used solely to represent phylogenetic groups is as follows: evolution is the single most powerful and general process underlying biological diversity. The major outcome of the evolutionary process is the production of an ever-branching phylogenetic tree, through descent with modification along the branches. This results in life being organized as a hierarchy of nested monophyletic groups. Since the most effective and natural classification systems are those that "capture" entities

resulting from processes generating the things being classified, the general biological classification system should be used to reflect the tree of life. Phylogenetic taxa will thus be "natural" in the sense of being the result of the evolutionary process.

This isn't to say that phylogeny is the only important organizing principle in biology, There are many ways of classifying organisms into a hierarchy, because of the many biological processes impinging on organisms. Many kinds of non-phylogenetic biological groupings are unquestionably useful for *special purposes* (e.g., "producers," "rain forests," "hummingbird pollinated plants," "bacteria"). However, it is generally agreed that there should be one consistent, *general-purpose*, reference system, for which the Linnaean hierarchy should be reserved. Phylogeny is the best criterion for the general purpose classification, both theoretically (the tree of life is the single universal outcome of the evolutionary process) and practically (phylogenetic relationship is the best criterion for summarizing known data about attributes of organisms and predicting unknown attributes). The other possible ways to classify can of course be used simultaneously, but should be regarded as special purpose classifications and clearly distinguished from phylogenetic formal taxa.

Important concepts in phylogenetic classification to revisit:

Monophyly

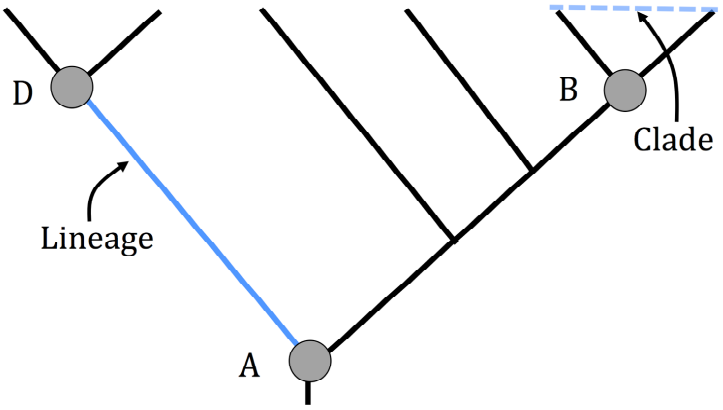
There are two different ways of defining monophyly: *synchronic* (i.e., "all and only descendants of a common ancestor") or *diachronic* (i.e., "an ancestor and all of its descendants"). Which is better?

Should the word "species" appear in the definition of monophyly? Does that matter? [It turns out it matters quite a bit as we'll see when we get to species concepts.]

The traditional cladistic concept of monophyly is itself in need of refinement in light of modern genomic data. Horizontal transfer (reticulation) is much more common in nature than realized 20 years ago. Despite being frequently presented as such, reticulation is not just a problem for the species level; clades at all levels can be subject to horizontal transfer. In the modern genomic world, because of the mounting evidence of horizontal gene transfer at all levels, monophyly can no longer mean monophyly of a group of organisms on every gene tree (as assumed by earlier generations of cladists, before there were data to the contrary) – we would have few to no monophyletic groups, at any level, in that strict sense. Rather, monophyly refers to an ensemble characteristic of organismic descent as discussed by Baum (2009). Monophyly refers to the preponderance of gene lineages making up a clade (using the clade-lineage distinction below). Gene lineages that don't match the pattern of descent shown by the majority of lineages need a different explanation (e.g., horizontal transfer or incomplete lineage sorting) than the majority. Note that this is analogous to the distinction people have made for a long time between homology and homoplasy; in fact, horizontal gene transfer is best viewed as a type of homoplasy.

Clade versus Lineage (see figure on next page for illustration).

They are not the same thing -- "clade" is a synchronic concept, a snapshot of a lineage at one time -- while a "lineage" is a diachronic concept, a series of replicators. This distinction is especially important when we get into species concepts in a later lecture.



The distinction between clades and lineages, showing the incompatibility of different views of species. A clade is a synchronic, monophyletic set of lineage-representatives, where monophyly is defined synchronically as "all and only descendants of a common ancestor" (represented by A' in this case). A lineage is a diachronic ancestor-descendant path (blue line up the left side of the tree), whereas a lineage segment is a part of a lineage that connects two nodes (A and D in this example). From Mishler & Wilkins 2018, *The hunting of the SNaRC*

Three major logical phases of a phylogenetic analysis:

(1) *Character analysis*

➔ the elements of a data matrix (i.e., OTUs, characters, and character-states) are assembled.

➔ this complex process involves considerable reciprocal illumination (since developing hypotheses of distinct, independent characters with discrete states goes hand in hand with developing hypotheses of homogeneous OTUs).

(2) *Cladistic analysis*

➔ the data matrix is translated into a phylogeny.

➔ reciprocal illumination is often involved here as well, since incongruence between characters or odd behavior of particular OTUs may lead to a return to phase 1 (a reexamination of OTUs, characters, and models) primarily to check for fit to assumptions of character analysis, i.e., OTUs should be homogeneous for the characters employed; characters should be discrete, heritable, and independent.

(3) *Classification and evolutionary studies*

➔ the phylogeny is translated into a classification, based on an assessment of the relative support for different clades.

➔ formal taxa (including species) are named here, on the basis of clear support for their existence as monophyletic cross-sections of a lineage, and for their utility in developing and discussing process theories.

➔ carry out comparative analyses...