Jan. 27, 2016**. Morphological data I: Character analysis; What is a data matrix? Homoplasy; Ontogeny & structure of plants & animals; The role of morphology**

> **Reading:** B.D. Mishler. 2005. The logic of the data matrix in phylogenetic analysis. In V.A. Albert (ed.), Parsimony, Phylogeny, and Genomics, pp. 57-70. Oxford University Press.

I. Homology: practice -- a.k.a. Character Analysis

This concept of homology discussed last lecture is clear in theory, but how do we recognize homology in practice? The best early codification of recognition criteria was that of Remane (Wiley, 1981): detailed similarity in position, quality of resemblance, and continuance through intermediate forms. Also, an important contribution of cladists has been the explicit formulation of a phylogenetic criterion:

**** a hypothesis of taxic homology of necessity is also a hypothesis for the existence of a monophyletic group ****

Therefore, congruence among all postulated homologies provides a test of any single character in question, which is the central epistemological advance of the cladistic approach. Individual hypotheses of putative homology are built up on a character-by-character basis, then a congruence test is applied to distinguish homologies (i.e., those apparent homologies that are congruent with other characters) from homoplasies (i.e., apparent homologies that are not congruent with the plurality of characters -- see following section).

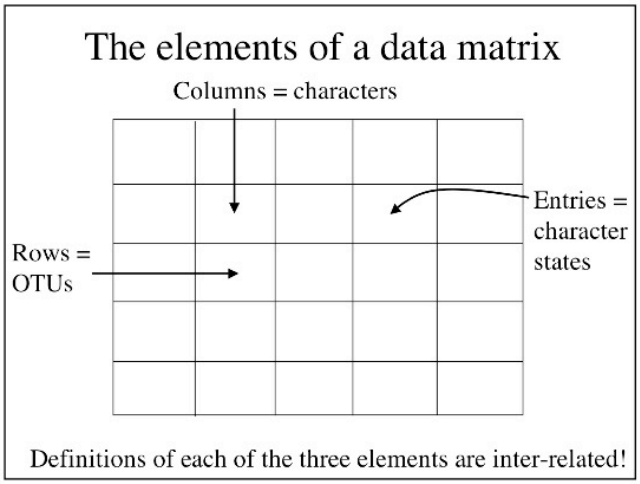
The central epistemological problem of systematic research is how to recognize, distinguish, and "define" taxonomic characters precisely, and choose the right ones for phylogenetic reconstruction at a particular level of interest.

Use the right tools for the job!

A. Introduction to the logic of the data matrix:

The full process of phylogenetic analysis inherently consists of three phases: first a data matrix is assembled, then a phylogenetic tree is inferred from that matrix, finally evolutionary analysis can be conducted using the tree. There is obviously some feedback between these phases, yet they remain logically distinct parts of the overall process. One could easily argue that the first phase of phylogenetic analysis is more important than the second phase; the tree is basically just a re-representation of the data matrix with no value added (Mishler, 2005).

Paradoxically, despite the logical preeminence of data matrix construction in phylogenetic analysis, by far the largest effort in phylogenetic theory has been directed at the second phase of analysis, the question of how to turn a data matrix into a tree. If we step back and take a hard look at the first phase, at stake are each of the logical elements of the data matrix: the **rows** (what are the terminal units or OTUs?), the **columns** (what are the characters?), and the **individual entries** (what are the character states?).



The elements of a data matrix (note the interlocking definitions):
OTU = group of semaphoronts that can't be subdivided given current character data
Character = an apparently homologous feature, independently varying among OTUs
Character-state = a discrete condition within a character, potentially a phylogenetic marker

B. What is an OTU?

These are represented by rows in the data matrix. People are usually cavalier about what their terminal branches represent. One often sees species or other taxon names, or even geographic designations of populations, attached to terminal branches of published trees without explanation. Larger-scale units *might* indeed be a well-justified OTU, but they need to be justified by preliminary analyses, never assumed a priori. Species or populations are never the fundamental things from which phylogenies are actually built. Not even individuals are the OTUs -- so what *is* the fundamental OTU?

As was carefully elaborated by Hennig (1966), the fundamental terminal entity in phylogenetics is the ***semaphoront***, an instantaneous time slice of an individual organism at some point in its ontogeny. A tube of extracted DNA and its associated museum voucher specimen, photos, sound recordings, or other data —a semaphoront— should be considered the ultimate unit of phylogenetic analysis. An OTU is an agglomeration of semaphoronts, that are not divisible by the characters currently known.

Hence, the interrelationship between the concept of OTU and character. [More later in the class when we cover species concepts.]

C. What is a Character?

Ontologically, taxonomic character (=putative taxic homology) is a piece of evidence for the existence of a monophyletic group. Epistemologically, a good taxonomic character is one that shows convincing **potential homology** across the OTU's being considered, and **shows greater variation among OTU's than within**. This variation must be **heritable and independent of other characters**, i.e., not genetically correlated with other characters in a specific evolutionary sense. Note that there are other meanings of "correlation", some of which (such as phylogenetic congruence) do not disqualify characters from counting as independent. Note also that this view of taxonomic characters requires that each be a **system of at least two discrete transformational homologs**, or *character states* (as discussed previously). Note that this is a restricted usage of the term "character," derived from the ontology of phylogenetic reconstruction. For other purposes, as in functional/evolutionary studies, numerical phenetic comparisons, or identification, less strict usages can be applied.

D. What is a character state?

The ontological view of taxonomic characters discussed above requires that each be a system of at least two discrete transformational homologs, or character states.

Epistemologically, the distinction of character states is a issue involving patterns of variation among OTUs. A reasonable statistical approach for quantitative data (Mishler & De Luna, 1991) is to use a standard ANOVA coupled with a multiple comparison test designed to discover which means are different from each other, and whether the means can be divided into groups that are significantly different from each other.

Character-state ordering and polarity will be discussed next lecture.

II. Homoplasy and traits

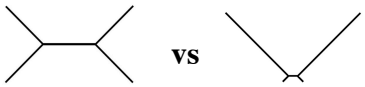
Homoplasy is similarity *not* due to historical continuity of information, a feature shared for one of several, distinctly different kinds of non-homologous reasons. Homoplasy can have various sources (see table below): "uncaused" (i.e., simple mistakes in gathering, interpreting, or compiling data, random matches between taxa, etc.) or "caused" (i.e., convergent evolution, reticulate evolutions, lineage sorting, developmental canalization, etc.). Homoplasy is viewed in systematics as an impediment to getting the correct phylogeny, but keep in mind that it can be studied in its own right. In fact, we'll see that much of the subject matter of this class is the study of homoplasy and its causes!

"Trait" is a more general term for a similarity among organisms, and the starting point for comparative phylogenetic studies. It includes both homology and homoplasy, and often the main goal of a comparative study is to sort out which is which. We will delve more into types of traits in later lectures.

A brief taxonomy of types of homoplasy:

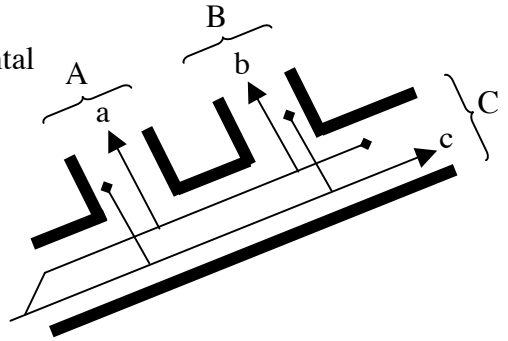
1. *Error* (e.g., mistakes in reading a gel, typographic errors, mislabeled specimens).

2. *Random matching over evolutionary time*. When a character has a limited number of states, non-homologous matches can occur -- this effect can cause biased reconstructions when the probability of change is very different in different lineages (the "long branch attraction" problem).



3. *Convergence*, due to natural selection in common environments.

4. *Parallelism*, perhaps due to shared developmental programs



4. *Reticulation* (e.g., hybrid speciation, introgression, horizontal gene transmission)

5. *Lineage sorting*, when different parts of the same genome have different branching histories due to differential extinction of polymorphisms.

III. Ontogeny

The last frontier in our understanding of biological forms is an understanding of their developmental origins. Much of the ultimate control of form resides in the genome, yet much also resides in the environment (at levels from the internal cellular environment to the external habitat). The highly interactive and complex nature of developmental processes make it impractical to deduce phenotype from genotype based on first principles. The phenotype is an emergent property and its origin can be studied most efficiently by backtracking from the phenotype itself to its structural, physiological, developmental and genetic causes. Development and morphology will remain a rich source of information for systematics and for evolutionary biology. We'll return to ontogeny and phylogeny in a later lecture.

1. Uses of ontogeny in systematics:

- A source of new characters in juvenile phases
- A source of clarifying homologies and defining character states in mature phases
- A source for determining transformational homology among character states within a character (ordering)
- A source for hypothesizing evolutionary directionality among character states within a character (polarization)

2. Development in animals:

1. Mosaic development (also called determinate)
   - Instructions are (mostly) built in and thus largely independent of environment.
   - Common in protostomes (worms, flies, etc.)
2. Regulative development (also called indeterminate)
   - Instructions come (mostly) from outside, highly dependent on cellular environment
   - Common in deuterostomes (vertebrates, sea urchins)

3. Development in plants (differences with animal development):

- Plants have modular growth, at several hierarchical levels
- Plants grow from an apical meristem (or single apical cell)
- Plant cells don't move (rigid cell wall)
- Plants do not have a segregated germ line

IV. Concluding thoughts on the roles of morphology

Why morphology in this day and age? Does it have any role? Some workers (e.g., Scotland, Olmstead, and Bennett, 2003) have argued that the active use of morphology in phylogenetic reconstruction is dead, and that phylogenies should be based solely on molecular data, relegating morphological characters to be passively mapped onto phylogenies later.

Such an argument unwisely downplays the value of morphological characters (as being too subjectively defined and evolutionarily plastic) while conveniently forgetting that molecular characters are subject to the same uncertainties about homology and character analysis, and may be quite homoplastic as well. It is much better to take a hard look at the advantages and disadvantages of each kind of data, according to the criteria we discussed last time. First, let's start with the roles that morphological characters can play, and do the same for molecular data later (next week).

Brent's Top Ten reasons to include morphological characters in phylogenetics:

10. Their greater complexity may allow better homology assessments. Unlike DNA sequences, which are often one-dimensional strings (unless you have secondary structure), morphology is complex and three-dimensional, plus has ontogeny.

9. They have many potential character states. As we will see later in the semester, an important parameter determining whether your data might be subject to "long-branch attraction" problems is the number of potential character states. False reconstructions are only a problem when parallel changes to the same character state happen, a phenomenon that is most frequent with binary data and rare with many available states.

8. **Data can be gathered from *many* specimens, cheaply and quickly.** A systematist can base their conclusions on samples from thousands of semaphoronts.

7. We need to be able to identify lineages easily in the field. Morphological apomorphies are easier to apply in field keys and in photo IDs guides.

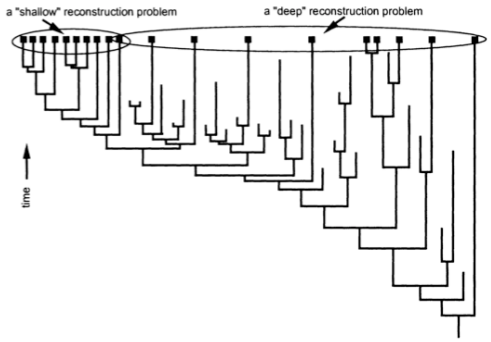
6. Discovering morphological apomorphies. We need to have a real analysis to show what the apomorphies at a particular level are. It is not rigorous to inspect a purely molecular tree and hang morphological characters onto branches intuitively.

5. Morphology gives you another independent data set, distinct from your organellar and nuclear genes. Comparing the topology of morphological datasets to those derived from specific genes can help you discover reticulation, lineage sorting, etc.

4. Morphological characters might actually help you get the best-supported answer! Even in cases where the topology of the total evidence tree is the same as with the molecules alone, support values such as bootstrap values often go up. And sometimes, the total-evidence topology has novel, highly-supported branches, synergistically supported by the combined data.

3. Episodic patterns of change. Despite common misconceptions to the contrary, clock-like markers are actually undesirable for reconstructing deep, short branches. Such markers continue to click along, changing at a regular rate until all the signal marking the deep branch is gone. The best marker for such deep branches is like the clock on the *Titanic* -- ticks once and stops forever. Slow change with long periods of stasis works best for these cases, i.e., the pattern shown by some morphological and anatomical features.

2. Better sampling of the tree of life. As we'll study later, good sampling is extremely important for reconstructing the correct tree. We need to break down those long branches. 99%+ of the lineages that have existed on the tree of life are extinct, and the only feasible way to get information about them is by adding fossils, which in turn requires morphology.



1. Studies of molecular clocks and dating of lineages. In order to include fossils, we must have morphological characters in the matrix, and therefore optimized to the cladogram. The fossils do not come with a taxon ID in the fossil record; they just come with some morphological characters. The fossil must therefore be attached to the cladogram based on its characters, then (and only then) can we infer that its sister group is at least as old as its age.

<u>The Bottom line:</u> you have to have a rigorous morphological character matrix to achieve most of the goals of phylogenetics, including incorporating information from fossils in phylogenetics, getting the tree right, and interpreting character evolution rigorously.