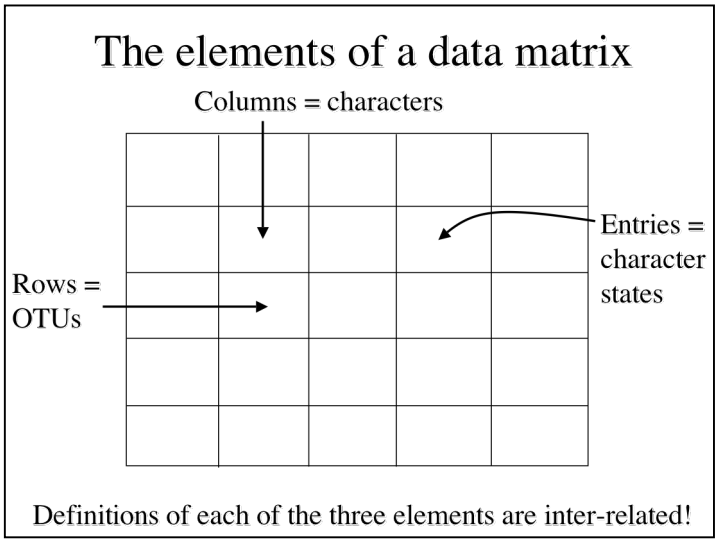
*Integrative Biology 200A* "PRINCIPLES OF PHYLOGENETICS" Spring 2008
University of California, Berkeley B.D. Mishler
Jan. 31, 2008**. Character Analysis**

**A. Introduction to the logic of the data matrix:**

The process of phylogenetic analysis per se inherently consists of two phases: first a data matrix is assembled, then a phylogenetic tree is inferred from that matrix. There is obviously some feedback between these two phases, yet they remain logically distinct parts of the overall process. One could easily argue that the first phase of phylogenetic analysis is the most important; the tree is basically just a re-representation of the data matrix with no value added.




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. 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. Taxa 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, data, etc. —a semaphoront— should be considered the ultimate unit. An OTU can then be best defined as: *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.]

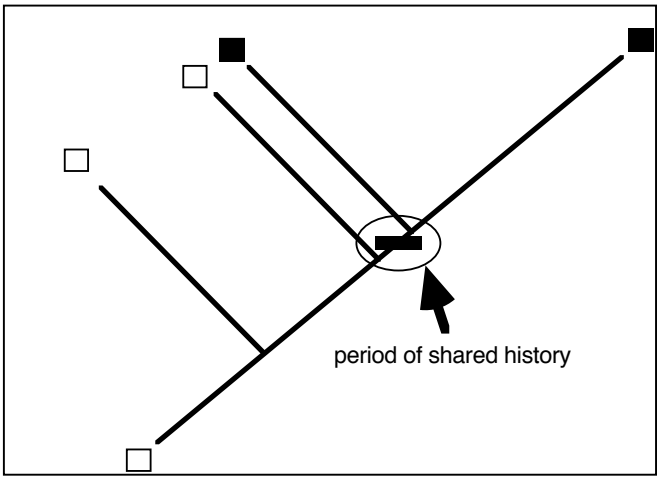
**C. What is a Character?**

The basic ontological stance taken here is:

***A taxonomic character (=putative taxic homology) is a piece of evidence for the existence of a monophyletic group.***

The central epistemological problem of systematic research is thus how to recognize, distinguish, and "define" taxonomic characters precisely. The selection of characters can be viewed as a sort of *a priori* weighting (Neff, 1986), but realize that this is quite a different issue than *a posteriori* weighting (more on that in a later lecture).

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* (more below). Note that this is a restricted usage of the term "character," derived from the ontology of phylogenetic systematics. For other purposes, as in functional/evolutionary studies, numerical phenetic comparisons, or identification, less strict usages are 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. How can we divide a "quantitative" character into states? There are two extremes of opinion: gap coding (e.g. Archie, 1985, Syst. Zool. 34: 326-345) vs. elimination (= "range coding" -- e.g. Pimentel and Riggins, 1987). Both are too extreme.

A reasonable intermediate approach can be outlined as follows (see also Mishler & De Luna, 1991): Given three OTU's (A,B,C) already tentatively distinguished on the basis of other characters, and the concept of character defended above, the question is: does a new quantitative character (one for which A, B & C have different means, but overlapping ranges, with A the largest, C the smallest, and B intermediate) provide evidence for the existence of monophyletic groups within this collection of OTU's? In more precise terms: is there a statistically significant association of the quantitative feature with the a priori discrete groups? An even more precise way to form the question is: can we reject the null hypothesis that the means of each OTU for this feature came from the same underlying parametric distribution? This is a standard type of question in science, and there is a generally applicable body of statistical methods known as analysis of variance to deal with such questions. In particular for our purposes we are interested

in the various multiple comparison tests designed to discover "which pairs of means are different from each other, and whether the means can be divided into groups that are significantly different from each other." (Sokal and Rohlf). [see handout on *Tortula*]

An all too common problem, however, with the results of multiple comparison tests is a situation as follows:

| | | | |
|---|---|---|---|
| statistically homogeneous subsets: | A | B | C |

What to do? It may be that B is heterogenous, and if broken up into two OTU's might resolve the situation. This is not valid if the overlapping variation occurs within individual organisms or interbreeding populations. In the latter case, the situation must represent one of the five resolutions shown below, and the apparent statistical overlap is an artifact of inadequate sampling:

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | B | C | A | B | C | A | B | C | A | B | C | A | C | B |

Given this situation, there is only one valid course of action: sample more representatives of the OTU's. If further sampling is not feasible, then the only valid coding for taxon B is "?", unknown. This represents an honest assessment of the current evidential meaning of the character: we don't know if B should be considered in the same character state as A, or as C, or in its own intermediate state. All we know is that A and C are in different states. To code the character otherwise is to step beyond the evidence at hand. Particularly, to code B as having an intermediate state is invalid. That represents an arbitrary choice among several possible resolutions.

**E. The problem of "polymorphism":**

-When a character that varies discretely elsewhere in the study group shows two different states within some individual OTU, you've got problems. Several different solutions are possible, depending on the nature of the situation:

(1) If the OTU appears phylogenetically heterogeneous, it should be divided up for purposes of analysis.

(2) If the variation occurs within individuals, then it might be necessary to code the OTU as unknown for the character, as detailed above.

(3) In the special case of character states segregating within interbreeding populations (as in electrophoretic alleles), it may be best to code the polymorphism as an intermediate state between the two fixed states.

**F. Summary of the practical process of character analysis**.

First:

(1) study previous literature on group; *all* previously suggested characters must be dealt with somehow (either by using them directly, modifying them, or eliminating them with just cause).

(2) scan through all available specimens of study group, without much attention to previous classifications; note variable features; “gestalt,” "intuition;" come up with new potential characters.

*These two steps produce your list of candidate characters*

Then:

(3) Examine this list of potential characters; describe carefully and if possible quantify; take measurements, do statistical analysis to text for discrete states.

(4) examine ontogeny; provides information on characters, character correlations, transformational homology, polarity.

(5) carry out growth experiments; provides information on character correlations, heritability.

The net result:

(6) a subset of the potential characters will survive all of above a priori tests, these are the taxonomic characters entered into the data matrix for the next phase, cladistic analysis.

Note that these "final" hypotheses of taxic homology are "final" only in a local sense; another round of character analysis can (and usually does) follow a preliminary cladistic analysis.

**G. Conclusion**

All characters have a limited window of utility, since they are defined with respect to a particular branching question at a particular level. Thus, characters need to be re-analyzed for each level of analysis -- no "automatic" approaches are justified. This gives rise to a problem of scaling and considerations of how to concatenate analyses at different scales.

How to scale up? We will deal with this issue in detail once we tackle methods for tree-building, but for now let's consider some alternatives initially:

**1. consensus coding** -- for a heterogeneous terminal group (ad hoc?)
**2. exemplars** -- choice of representative extant taxa ("basal"?), often combined with **supertrees**
**3. a "supermatrix" approach** -- where all semaphoronts are included in all analyses.
**4. compartmentalization** -- by analogy with a water-tight compartment on a ship; homoplasy isn't allowed in or out; cut data sets down to manageable size and allow use of more information in an analysis through improved homology assessments within compartments (e.g., character-state divisions in morphology; alignments in molecular data); thus, suppress the effect of spurious homoplasy. Procedures in compartmentalization:

i. global analysis, determine best supported clades (= compartments)
ii. local analyses within compartments, often with augmented data sets
iii. return to global analyses, either:
(a) with compartments constrained to local topology (for smaller data sets); or
(b) with compartments represented by a single HTU -- the inferred archetype