"PRINCIPLES OF PHYLOGENETICS: ECOLOGY AND EVOLUTION"

Integrative Biology 200 Spring 2014

University of California, Berkeley B.D. Mishler

Jan. 27, 2014**. Phylogenetic reconstruction in a nutshell: homology, characters, and traits**

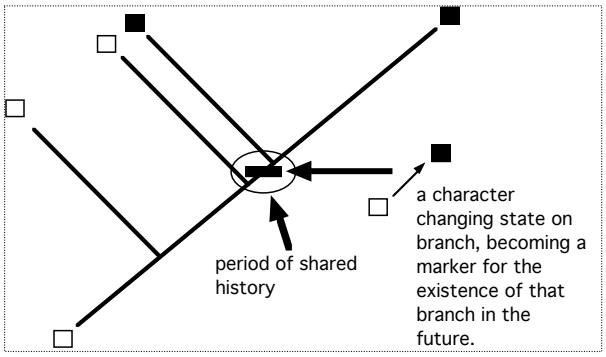
Reading: B.D. Mishler. 2005. The logic of the data matrix in phylogenetic analysis. In V.A. Albert (ed.), Parsimony, Phylogeny, and Genomics, pp. 57-70. Oxford University Press.

I. Introduction

Genealogical relationships themselves are invisible, so how can we know them? Is there an objective, logically sound method by which one can reconstruct the tree of life? Recent advances in theories and methods for phylogenetic reconstruction, along with copious new data from the molecular level, have made possible a new scientific understanding of the relationships of organisms. This understanding of relationships has lead in turn to improved taxonomic classifications as well as the subject matter of this class: comparative methods for testing biogeographic, ecological, behavioral, and other functional hypotheses.

II. The Hennig Principle

The fundamental idea driving recent advances in phylogenetics is known as the Hennig Principle, and is as elegant and fundamental in its way as was Darwin's principle of natural selection. It is indeed simple, yet profound in its implications. It is based on the idea of homology (more on homology in the next section).



Hennig's seminal contribution was to note that in a system evolving via descent with modification and splitting of lineages, characters that changed state along a particular lineage can serve to indicate the prior existence of that lineage, even after further splitting occurs. The "Hennig Principle" follows from this: homologous similarities among organisms come in two basic kinds, synapomorphies due to immediate shared ancestry (i.e., a common ancestor at a specific phylogenetic level), and symplesiomorphies due to more distant ancestry (see figure on next page). Only the former are useful for reconstructing the relative order of branching events in phylogeny -- "special similarities" (synapomorphies) are the key to reconstructing truly natural relationships of organisms, rather than overall similarity (which is an incoherent mixture of synapomorphy, symplesiomorphy, and non-homology).

Classifications are applied to the resulting branching diagram (cladogram). A corollary of the Hennig Principle is that classification should reflect reconstructed branching order; only monophyletic groups should be formally named. A strictly monophyletic group is one that *contains all and only descendents of a common ancestor*. A paraphyletic group is one the excludes some of the descendents of the common ancestor. We will return to deal with the ramifications of this approach to classification next week and throughout the course.

This elegant correspondence between synapomorphy, homology, and monophyly is the basis of the cladistic revolution in systematics.

III. Homology: theory

Homology is of the most important concepts in biology, but also one of the most controversial. What does it mean to say that two organisms share the *same* characteristic? The modern concept refers to a continuity of information from ancestor to descendant (not identity!!). Homology is defined as *a feature shared by two entities (e.g., organisms, genes) because of descent from a common ancestor that had that feature*. There are thus two types of homology that we are concerned with here: phylogenetic homology, which is the same character state in two different lineages at one time-slice (i.e., synapomorphy); and transformational homology, which is the relationship through time in one lineage between character states (i.e., the relationship between an apomorphy and its plesiomorphy). Specific hypotheses of transformational homology among character states are called transformation series.

A. Types of homology
- --Iterative Homology (within one organism), e.g., Serial Homology or Paralogy in molecular data
- --Phylogenetic Homology (between organisms)
  - --Taxic (= synapomorphy)
  - --Transformational (plesiomorphy -> apomorphy)

B. How do we recognize homology?
- -- Remane's criteria (detailed similarity in position and quality of resemblance)
- -- Congruence test (a recently formulated, explicitly phylogenetic criterion)

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

This concept of homology 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).

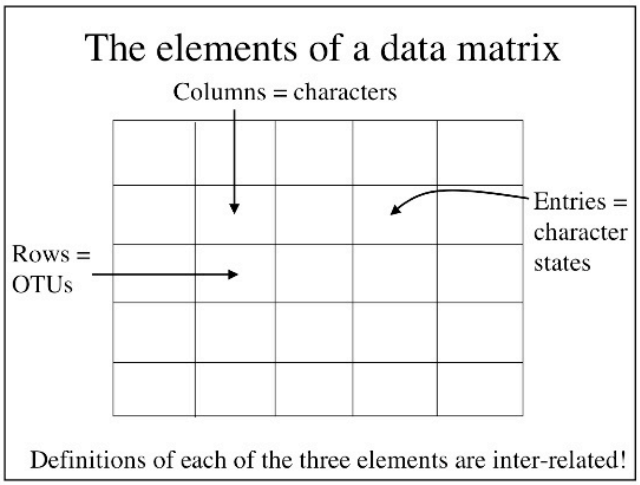
The central epistemological problem of systematic research is how to recognize, distinguish, and "define" taxonomic characters precisely, and choose the right ones for phylogenetic reconstruction at a particular level of interest.

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. Specific hypotheses of transformational homology among character states are called *transformation series*. "Ordering" refers to the specification of character state "adjacency" without any implied directionality (N.B., not the same as polarity). Such specifications are best made from studies of ontogeny, where one can often directly observe transformations between character states. Sometimes (and perhaps reasonably), these specifications are made from observations of "morphoclines" in adults. In many cases, however (e.g., alternative bases at a homologous site in molecular sequence data), no reasonable evidence exists for ordering, and states are best left "unordered." When in doubt, it is best to err on the conservative side and leave characters unordered, but note that potential phylogenetic information is always lost when doing so. It is also possible, using the "step matrix" function in PAUP, to code characters as partially ordered.

Character-state polarity. Determining which character state of a transformation series is plesiomorphic is called the problem of evolutionary polarity. Several methods have been advocated, but only three are widely used: (1) paleontology -- the state occurring earliest in the fossil record is considered plesiomorphic; (2) ontogeny -- the state occurring earliest in development is considered plesiomorphic; (3) outgroup -- the state occurring outside the study group is considered plesiomorphic. All three have potential problems, but the last is the one most widely recommended.

V. Homoplasy and traits

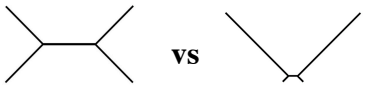
Homoplasy is similarity *not* due to historical continuity of information, a feature shared for one of several, distinctly different kinds of non-homologous reasons. Homoplasy can have various sources (see table below): "uncaused" (i.e., simple mistakes in gathering, interpreting, or compiling data, 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.