Integrative Biology 200A "PRINCIPLES OF PHYLOGENETICS" Spring 2010
University of California, Berkeley B.D. Mishler

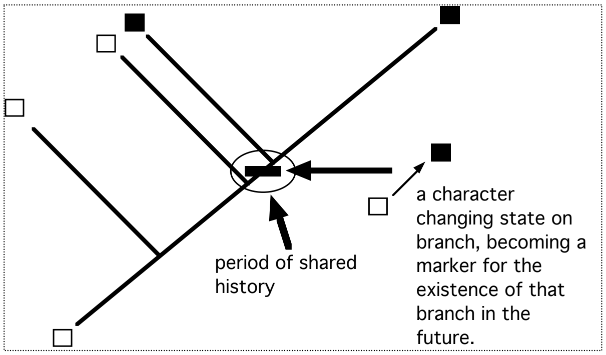
Jan. 26, 2010. The Hennig Principle: homology; synapomorphy; rooting issues; character analysis -- what is a data matrix?

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 a wealth of 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, one of the most important concepts in systematics, but also one of the most controversial. What does it mean to say that two organisms share the same characteristic? The modern concept is based on evidence for historical continuity of information; homology would then be defined as *a feature shared by two organisms because of descent from a common ancestor that had that feature* (more on homology below).


a character changing state on branch, becoming a marker for the existence of that branch in the future.
period of shared history

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. 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 later in the course.

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

III. Homology

We must pay close attention to both ontology and epistemology, and the feedback relationship between the two: A given method makes sense only if the world really is a certain way, yet the view we have of how the world is organized is dependent on the methods we have used. For example, if species on earth are related genealogically and evolution is mainly by descent with modification (in a primarily diverging mode), then the Hennig Principle is the best method for reconstructing the history of life. Yet, the discovery of hierarchically nested characters is the best evidence we have on how evolution has occurred.

When are two things the same?

These concerns are relevant to characters as well; the mere act of stating that two things are the same, or parts of two things are the same, is loaded with a (perhaps subconscious but nonetheless real) complex theoretical framework.

"Homology" -- One of the most important concepts in systematics, but also one of the most controversial. History:

--classical, pre-evolutionary views (Cuvier, Owen)
--nominalistic views (many botanists, pheneticists)
--developmental views.
--evolutionary views: historical connectedness.
--synapomorphy (Patterson, Stevens)
--historical continuity of information (Van Valen, Roth)**

Our ontology. A continuity of information from ancestor to descendant (not identity!!). *A homology is a similarity due to historical continuity of information, a feature shared because of descent from a common ancestral 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)

Our epistemology. This concept is clear in theory, but how do we recognize homology? 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 <u>homoplasies</u> (i.e., apparent homologies that are not congruent with the plurality of characters).

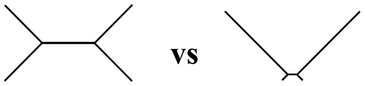
Is this circular? A quick digression into general concerns in the philosophy of science; reciprocal illumination.

IV. Homoplasy

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: "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 mine 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! Here is 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).

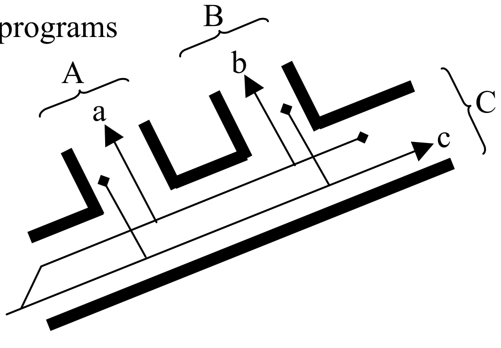


3. *Convergence*, due to natural selection in common environments.

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.



V. Character Analysis

Now that we have the basic ontology in mind, how best to proceed with empirical research? The basic stance taken here is:

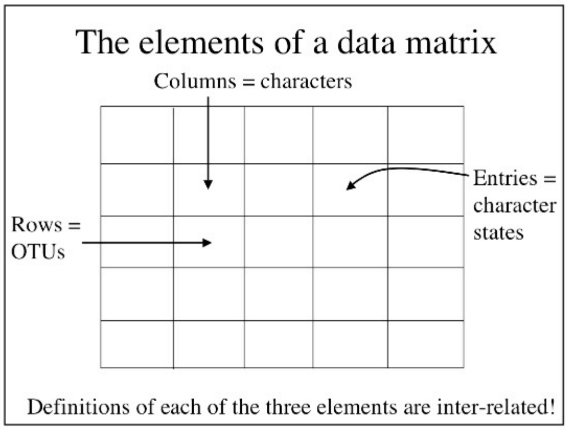
A taxonomic character (=putative taxic homology) is a piece of evidence for the existence of a monophyletic group.

The central epistemological problem of systematic research is 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
Columns = characters
Entries = character states
Rows = OTUs
Definitions of each of the three elements are inter-related!

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, sound recordings, or other data —a semaphoront— should be considered the ultimate unit. 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.

Polymorphism is when a character that varies discretely elsewhere in the study group shows two different states within some individual OTU. 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.

(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.

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.

An alternative, commonly applied approach to polarizing the characters before an analysis is to first construct the topology of the tree as an unrooted network, and then "pull it down" into a tree shape in one of two ways:

(1) By bending at the point where the outgroup joins the ingroup: Outgroup rooting.
(2) By seeing where an ancestral vector of hypothesized character states would attach: Lundberg rooting.

The correct way to use these approaches will be briefly discussed here, along with some cautions, but we'll need to return to this important issue later in the semester.

VI. 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.

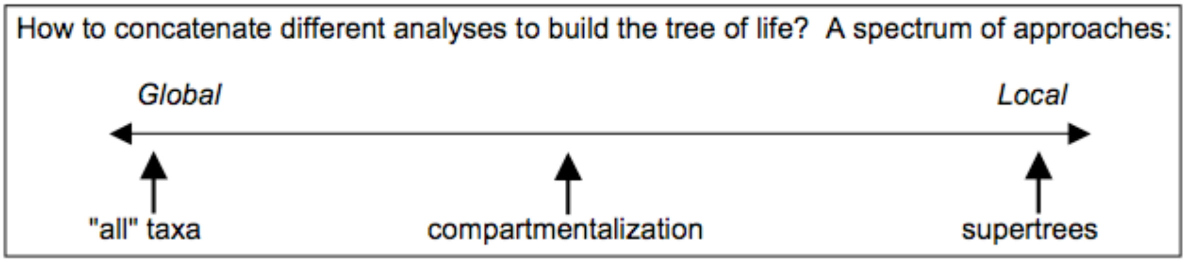
VII. Scaling up to the tree of life.

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 more 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 and all characters are included.
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 <u>within</u> 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


How to concatenate different analyses to build the tree of life? A spectrum of approaches:

VIII. Conclusion: General properties of good phylogenetic markers

As argued above, the fundamental advances leading to recent progress in systematics are in theory and method (about how to use *any* data for phylogenetic reconstruction). However, another advance that occurred at the same time, due to improvements in technology of molecular biology, was the availability of a large amount of new data at the molecular level. The incorporation of these new sources of data helped to reinvigorate systematics and fuel the recent quantum leap forward in knowledge of phylogenetic relationships. But to put the newer molecular characters in context, we need to examine the desired criteria for *any* character set to be useful as a marker at some level in the tree of life.

Remember, we are interested in finding a marker to infer the past existence of a period of shared history, without being mislead by what has happened since. Several things can derail this process of inference; thus, an ideal marker should have at least six properties (Wiley, 1981; Mishler & De Luna, 1991; Mishler 2000):

(1) *Complexity and comparability*. — We need to make a well-founded comparison between organisms for initial hypotheses of homology. Thus, a good taxonomic marker should be complex in structure and ideally in development as well, allowing a hypothesis of homology marking some particular shared branch even in the face of further change in that character.

(2) *Discrete states*. — A good taxonomic marker shows greater variation among than within OTU’s (Operational Taxonomic Units, a pragmatic grouping of organisms in an analysis that is homogeneous for the characters currently known). This view of taxonomic characters

requires that each be a system of at least two discrete transformational homologs, or **character states**.

(3) *Heritability*. — Variation in a character must be heritable—causally correlated between parent and offspring.

(4) *Independence*. — The probability of change in some particular character must be uncorrelated with any other character in a specific causal sense. One character must be free to change on the phylogeny without affecting the probability of another character changing. Note that there are other meanings of ”correlation”, some of which (such as phylogenetic congruence) do not disqualify characters from counting as independent—the dependence problem is confined to causal correlation at the time of change. There are a number of potential causes of character dependence, including developmental correlations linking morphological characters, secondary and tertiary protein and RNA structure linking different nucleotide positions, and natural selection which can cause suites of any kinds of character to covary as a block.

(5) *Low rate of change*. — One of the best ways to predict phylogenetic behavior of characters is by examining variation in the central parameter λ, defined as the **branch length** in terms of expected number of character changes per branch of a tree (Mishler, 1994). This parameter incorporates both rate of change per unit time and the length of time over which the branch existed (thus, a high λ can be due to either a high rate of change or a historically long branch). This parameter defines a ”window of informativeness” for that data. A very low value of λ indicates data with too few changes on each segment to allow all branches to be discovered; this would result in polytomies in reconstructions because of too little evidence. Too high a value of λ indicates data that are changing so frequently that problems arise with homoplasy through multiple changes in the same character. At best a high λ causes erasure of historical evidence for the existence of a branch, at worse it creates ”evidence” for false branches through parallel origins of the same state. Felsenstein (1978) showed that branch-length asymmetries within a tree can cause parsimony reconstructions to be inconsistent. That is, if the probability of a parallel change to the same state in each of two long branches is greater than the probability of a single change in a short-connecting branch, then the two long branches will tend to falsely ”attract” each other using a large number of characters.

(6) *Many possible character states*. — The effect of long-branch attraction is greatly moderated if a character has a number of distinct ways it can vary (Mishler, 1994). 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.

Properties of a good marker, as compared between molecules and morphology.

	molecules	morphology
1) COMPLEXITY AND COMPARABILITY	−	+
2) DISCRETE STATES	+	−
3) HERITABILITY	+	−
4) INDEPENDENCE	?	?
5) LOW RATE OF CHANGE (λ)	?	?
6) MANY POSSIBLE CHARACTER STATES	−	+

”Use all the characters that are fit to use”

One conclusions is clear: there is every reason to search carefully for good potential markers in all kinds of data. Thus, the tendency seen frequently these days to ignore morphological data in favor of molecular data should be avoided; all good characters should be sought and used.