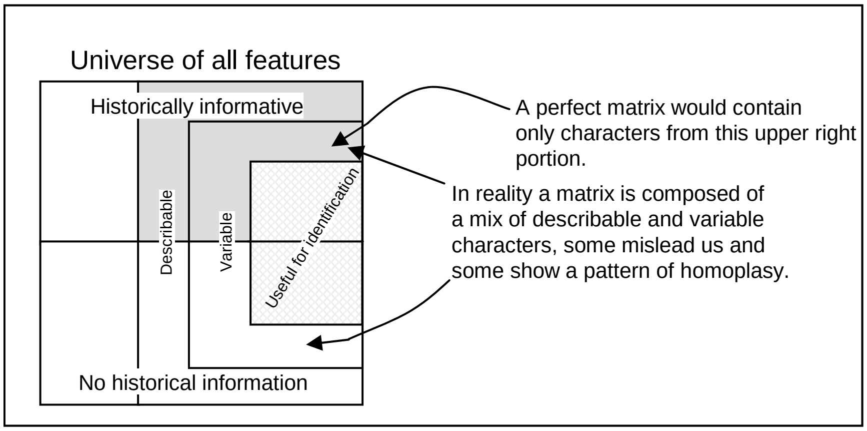
Integrative Biology 200A “PRINCIPLES OF PHYLOGENETICS” Spring 2008
University of California, Berkeley Kipling Will

12 Feb. **More on Characters and Character Coding**

To Reiterate and expand some ideas

- *Why are characters as or more important than the method of analysis*? Given a chosen algorithm and a set of coded characters with their assumed homology the set of trees is given. Different tree generating methods often result in the same or similar relationships. However, changing characters and character coding usually changes the actual meaning and interpretation of the evolution of the group (think about transformation series, hypothetical ancestors and branch lengths).



A perfect matrix would contain only characters from this upper right portion.

In reality a matrix is composed of a mix of describable and variable characters, some mislead us and some show a pattern of homoplasy.

- *Hennig’s Auxiliary Principle* - Assume homology in the absence of contrary evidence.
- *Conjectural homology* assessment (similarity, etc.) prior to cladistic analysis and *corroborated homology* assessment (homology [underlying process] = synapomorphy [pattern]) and homoplasy [pattern not due to a common process] after cladistic analysis.
- Characters = Transformation series= columns in matrix
- Character states= cell entries in matrix
- Characters states are the alternative forms of characters observed in the semaphorant.
- Character states are sub-grouping of characters, and characters are in turn states at a higher-level or dependent on states at a higher level.
- When the analysis results in corroboration of our initial hypotheses all superordinate characters remain unchallenged.

-When the analysis results in a pattern of homoplasy, i.e., our initial assumption of homology was wrong, the error could be at any level. We may decide that re-coding is appropriate.

- At some level no characters are absolutely independent. The level of non-independence varies along a gradient from negligible to covariance.
- Coding is an abstraction of observations.

    -Coding is a state-ordering process.

    -Methods require discrete states.

Informative vs. noninformative characters- In a parsimony based analysis various character state distributions may not provide grouping information. More on this when we discuss optimization. A character is uninformative if, according to its current transformation type, any possible dichotomous tree would require the same number of steps in the character. Under most model-based analyses no character is a priori uninformative. *Why?*

Polarity- *Direction of character change.* Distinguishing ancestral from derived or the idea of "polarizing" characters originated early in phylogenetics and was central to Hennig's (1966) phylogenetic method and reconstruction methods developed by Wagner (1961, 1969), etc. It was essential to identify primitive vs. derived character states prior to tree construction. Establishing character state polarity prior to analysis in many papers gave rise to a misconception that it is necessary to "polarize" characters. Determination of character polarity prior to cladistic analysis as it is now implemented is usually not desirable.

Some methods used to determining a priori polarity:
1. "Traditional" Outgroup comparison- Select an OTU or set of OTUs that is/are outside the in-group, but closely related to it, to be the outgroup(s) (best if it includes the sister group). Assume character states in outgroup are ancestral. Problems??

2. Hypothetical ancestor, sometimes as a "ground-plan" is constructed based on a composite idea of outgroup taxa and especially the notion that common equals primitive. Problems??

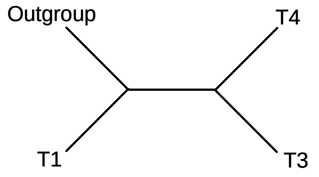
3. Embryological criteria- Application of modified Von Baer's law or as Haeckel proposed "ontogeny recapitulates phylogeny." General, primitive, ancestral characters appear in embryo before derived, e.g. Gills---->No Gills. Problems??

4. Paleontological criterion- Assume older fossils exhibit more ancestral characters. Problems??

5. Chorological progression- Species nearer the center of origin of the taxon have the primitive character states. Problems??

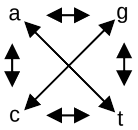
Current Outgroup analysis method- putative outgroup taxa are included in the analysis and the network is rooted between the ingroup and outgroup(s), and then "character polarity" is based on optimization of the character on a particular tree topology. This method avoids incorporation of preconceived bias into the analysis, allows testing of the monophyly of the ingroup (if more than one outgroup is employed). This method was first proposed by Farris (1972). Problems??

No polarity = 1↔0
(Outgroup (T1(T2, T3))) = (0(0(1,1))) This implies 0→1



Kinds of Characters and coding examples:

Binary character- two state character- 0,1
Multistate character- more than two states- 0,1,2...; ACGT
Nonadditive (=unordered) multistate character- No set character state adjacency. Same number of steps between any two states.



Additive (ordered) multistate character- Character with state-to-state adjacency specified such that in the analysis a violation of the ordering cost more steps.

0↔1↔2

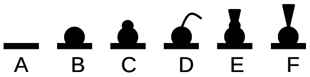
Can we justify setting character order (=make them additive or set adjacency)?

The same logic that is used to establish characters and character states and the hierarchical relationship of characters is the same at this level. You must be explicit about assumptions!

Binary and mixed coding of Additive multistate character- Hierarchies and other complex relationships between character states can be represented in the coding. [a.k.a Probable pathways models, character state trees]

Additive Binary coding

A					
B					
C					
D					
E					
F					



Unit coded characters- Binary characters that are grouped in sets to define a complex configuration. Also can represent reticulate transformation series.

An example from Liebherr (1998). [91-96. Pronotal marginal gutter broad, edge upturned (0,0,0,0,0,0); margin very broad, edge upturned (1,0,0,0,0,0); marginal gutter moderate, edge upturned (0,1,0,0,0,0); marginal gutter moderate, edge beaded (0,1,1,0,0,0); marginal gutter narrow, edge upturned (0,1,0,1,0,0); marginal gutter narrow, edge beaded (0,1,1,1,0,0); marginal gutter obsolete, marginal bead present (0,1,1,1,1,0); marginal gutter and bead absent (01,1,1,1,1). The derived states for characters 93-94 imply a reticulate character state transformation series….]

Using what you learn today to determine the transformation series above, where is the reticulation?

Mixed additive coding (a.k.a. Multistate hierarchic coding or linear nonredundant coding)

A			
B			
C			
D			
E			
F			

More explicit Evolutionary models: (Typically implemented by the software)
Irreversible characters- Multiple gains allowed, no losses
Dollo characters- Multiple losses allowed, multiple gains not.

Implicit in all of these are a character state **step-matrix, or cost matrix** (Sankoff, 1975), assigning costs to changes.

Unordered

$$\begin{array}{c|cccc} & 0 & 1 & 2 & 3 \\ 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 2 & 1 & 1 & 0 & 1 \\ 3 & 1 & 1 & 1 & 0 \end{array}$$

Ordered

$$\begin{array}{c|cccc} & 0 & 1 & 2 & 3 \\ 0 & 0 & 1 & 2 & 3 \\ 1 & 1 & 0 & 1 & 2 \\ 2 & 2 & 1 & 0 & 1 \\ 3 & 3 & 2 & 1 & 0 \end{array}$$

Irreversible

$$\begin{array}{c|cccc} & 0 & 1 & 2 & 3 \\ 0 & 0 & 1 & 2 & 3 \\ 1 & \infty & 0 & 1 & 2 \\ 2 & \infty & \infty & 0 & 1 \\ 3 & \infty & \infty & \infty & 0 \end{array}$$

-Step matrices can be used for any number of transformation or weighting schemes, even asymmetrical ones and ones with non-zero diagonals, e.g., transversions cost more:

$$\begin{array}{c|cccc} & A & C & G & T \\ A & - & 2 & 1 & 2 \\ C & 2 & - & 2 & 1 \\ G & 1 & 2 & - & 2 \\ T & 2 & 1 & 2 & - \end{array}$$

-Values in the step matrix can be steps or probabilities or any relative measure.

-**The triangle inequality**. In defining matrix of state-state distances, can violate a fundamental property of distances called the "triangle inequality." Triangles in Euclidean space have property that the length of one side is always less than sum of lengths of other two sides— or, shortest distance between two points is straight line. (Can't have indirect route that provides shortcut) programs like MacClade will allow the inequality exists but warn with a caution message; PAUP will auto adjust matrix to satisfy the triangle.

Weighting vs. Cost- A character has a cost, i.e. the total number of steps or state transformations on a given topology adjusted by costs in the step matrix. A character also has a weight, i.e. a factor applied to any change in the character states. This acts the same as having many characters with the exact same state distribution in the matrix. Although the vast majority of people agree that all characters are not equally “good”, equal-weights (which is a kind of weighting) is most commonly used.

----Scoring “missing” data----

Notation typically used:
? [unknown or not applicable]
- [gap in sequences, can be fifth character OR equal “?”]
*[complete polymorphism, all states observed]
$[subset polymorphism, e.g. only states 0 and 1 observed for taxon for a character with states 0,1,2]

Missing data entries are used when
1. state is not known but the character presumably exists in the semaphorant
2. character is not applicable as the structure does not exist in the semaphorant
3. the character is polymorphic in the OTU

-In number 1 this is the best treatment available as it represents the fact that we are ignorant.

-Number 2 is problematic. Maddison (1993) presents the classic red-tail/blue-tail/no tail example showing that dividing tail characteristics into two characters, tail color and tail presence, can lead to an interaction between distant clades that may result in a failure to consider some reasonable resolutions. Coding the states as a single character, e.g. gaps as a fifth state in sequence data may create an undesirable equivalence between very different types of change. In the tail example, coded as a single multistate character it equates a change in tail color to gain/loss of a tail.

-Number 3 is a case discussed by Nixon & Davis (1991), and it may be handled by decomposition of polymorphic terminal into monomorphic component parts, inferring ancestral states or leaving the terminal as “missing”.

Depends on how well you know the OTU. If the OTU can be reasonably assumed as monophyletic and the characters really occur in all combinations then scoring polymorphic as missing is correct. If your terminal is a large group, e.g., Insecta, and you have multiple cells with polymorphic states, you may have interaction among characters that result in implausible character state combinations.

If possible monomorphic terminals are better, as cells with missing values for polymorphic OTUs always underestimate tree length, characters are not fully contributing to the resolution of the tree and ancestral states cannot be assigned.