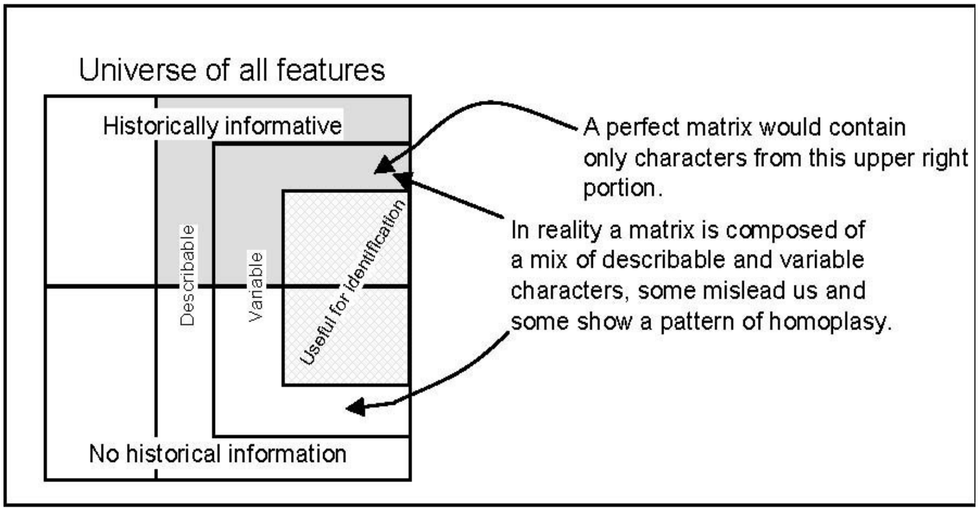
Integrative Biology 200 "PRINCIPLES OF PHYLOGENETICS" Spring 2016
University of California, Berkeley K.W. Will

On Characters, where they come from and Character Coding

Why are characters as or more critical than the chosen method of analysis? Given an algorithm or tree searching/building method and a set of coded characters (primary homologies) the trees are largely 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).


Universe of all features
Historically informative
Describable
Variable
Useful for identification
No historical information
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.

In search of the perfect character matrix: If possible, we would have only discretely describable, appropriately variable, independent, historically informative characters in our matrix. All of these are difficult or impossible (in an absolute sense) to achieve.

1. Discretely describable: There is a degree of subjectivity and impact of technology on what we can observer and level of detail we can describe.
2. Appropriately variable: Obvious ends of the spectrum, uniform – all unique, are easily eliminated, but in between we can only estimate what a reasonable amount of change is.
3. Quasi-independent: In some sense all characters are interdependent as they are together in a semaphorant and have shared history. But, there is a range from fully dependent to effectively unlinked.
4. Historically informative: The process of character analysis is the first hurdle and then corroboration is the test for this.

The matrix we end up with isn't perfect but it isn't a random sampling.

Homology: From observations, to characters, to corroboration.

Conjectural homology or *primary homology:* Our initial observation and assessment based on similarity, position, development, and test for conjunction done prior to cladistic analysis. This compels us to postulate homology. Tests of similarity (in the broad sense) are not tests of homology, nor do they result in patterns of homoplasy. "*Hypotheses of homology are conjectures whose source is immaterial to their status*" (Patterson 1982, see also de Pinna 1991).

Hennig's Auxiliary Principle: Assume homology in the absence of contrary evidence.

- Characters = a collection of features of semaphorants = Transformation series= columns in matrix
- Character states= cell entries in matrix. States are an alternative expression of the character as observed in the semaphorant.
- Character states are nested in characters, and characters at the level of interest are in turn states at a higher-level or dependent on homology at a higher level.

Corroborated homology: Our primary homology statements of characters and their states then are brought together in our analysis and are corroborated or refuted, but never proven. Cladistic methods are fundamentally trying to maximize the corroboration of our primary homology statements. If our primary homology assessment is not refuted we can think of the homology in terms of the underlying process (diachronic transformation) or equal to synapomorphy, which is in a sense a pattern term (synchronic or instantaneous). When the analysis results in corroboration of our initial hypotheses all superordinate characters remain unchallenged.

When our primary homology statements are refuted at the level we are testing, this results in a pattern of homoplasy. Homoplasy is a pattern term and is not necessarily due to a common process. The processes invoked are ad hoc and explain a pattern of homoplasy.

Coding: an abstraction of observations

- Coding is a state-ordering process.
- Most methods require discrete states, though there are methods for continuous characters.

Polarity- Direction of character change. Distinguishing ancestral from derived states or the idea of "polarizing" characters originated early in phylogenetics and was central to Hennig's (1966) phylogenetic method and reconstruction methods (See Wagner 1961; Farris 1970), 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 is now 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.

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.

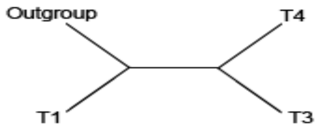
3. Embryological criteria- Application of modified Von Baer's law or as Haeckel proposed "ontogeny recapitulates phylogeny." General, primitive, ancestral characters appear in embryo before derived, e.g. Gills---->No Gills.

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

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

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). For more about this method see Nixon and Carpenter (1993).

No polarity = 1↔0
(Outgroup (T1(T3, T4))) = (0(0(1,1))) This implies 0→1



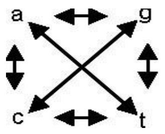
Kinds of Characters and coding examples.

Keep in mind that the trees and graphs of character states (probable pathways of transformation) are not the same thing as the character history on a phylogenetic tree.

Binary character- two state character- 0,1

Multistate character- more than two states- 0,1,2...; ACGT

Unordered ("Fitch parsimony") multistate character with no set character state adjacency same number of steps between any two states.

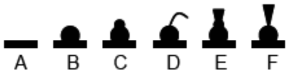


Ordered Additive ("Wagner or Farris parsimony") multistate character- Character with state-to-state adjacency specified such that in the analysis a violation of the ordering cost more steps. Ordering is not polarizing.

0↔1↔2

Can we justify setting character order (i.e 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. Alternatively, some view this as specific models of evolution. However, it is done you must be explicit about assumptions!

Binary and mixed coding of nonadditive 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 (graphs)].
-draw character state tree
-make one state the root (this is arbitrary, makes coding easier, but doesn't change costs)
-code group membership using 0 or 1 for all components including the terminal ones
-non-group states are 0



Binary coding and Linear 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 and Zimmerman (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)."]
Using what you learn today to determine the transformation series above, where is the reticulation?

Mixed coding- (a.k.a. Multistate hierarchic coding or linear nonredundant coding)
-draw character state tree
-make one state the root (this is arbitrary, makes coding easier, but doesn't change costs)
-code group membership for all states from the root to one terminal using multi-state numbers (0,1,2...)
-code any states not on the path as the value of the subtending state
-code each terminal not on path

A			
B			
C			
D			
E			
F			

More explicit Evolutionary models: (Typically implemented by the software)
Irreversible characters ("Camin-Sokal parsimony")- Multiple gains allowed, no losses
Dollo parsimony- 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

	0	1	2	3
0|	0	1	1	1
1|	1	0	1	1
2|	1	1	0	1
3|	1	1	1	0

Ordered

	0	1	2	3
0|	0	1	2	3
1|	1	0	1	2
2|	2	1	0	1
3|	3	2	1	0

Irreversible

	0	1	2	3
0|	0	1	2	3
1|	∞	0	1	2
2|	∞	∞	0	1
3|	∞	∞	∞	0

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

	A	C	G	T
A|	-	2	1	2
C|	2	-	2	1
G|	1	2	-	2
T|	2	1	2	-

Weighting vs. Cost- A character has a cost, i.e. the total number of steps or state transformations on a given topology adjusted by modifiers 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 [not scoped] but the character presumably exists in the semaphorant .This is the best treatment as it represents the fact that we are ignorant.

2. Character is not applicable as the structure does not exist in the semaphorant. This may be 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.

3. The character is polymorphic (a case discussed by Nixon and 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.

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.

Cited and recommended papers:

de Pinna MCC (1991) Concepts and tests of homology in the cladistic paradigm. Cladistics 7, 367-394.

Ferris, J.S. 1970. Methods for computing Wagner trees. Systematics Zoology. 19, 83-92.

Hennig, W. (1966). “Phylogenetic Systematics”. University of Illinois Press, Urbana, Illinois.

Liebherr, J.K. and E.C. Zimmerman. 1998. Cladistic analysis, phylogeny and biogeography of the Hawaiian Platynini (Coleoptera: Carabidae). Systematic Entomology 23,137-172.

Maddison, W. P. (1993). Missing data versus missing characters in phylogenetic analysis. Systematic Biology. 42, 576–581.

Nixon, K.C. and J.M. Carpenter. 1993. On Outgroups. Cladistics. 9, 413-426.

Nixon, K.C. and J.I. Davis. 2005. Polymorphic taxa, missing values and cladistics analysis. Cladistics. 7, 233-241.

Patterson, C. (1982). Morphological characters and homology. In “Problems of Phylogenetic Reconstruction” (K. A. Joysey, and A.E. Friday, Eds), pp. 21–74. Academic Press, London and New York.

Wagner, W. H. 1961. Problems in the classification of ferns. Recent advances in botany. Univ. Toronto P ress, Toronto. pp841-844.