Integrative Biology 200A "PRINCIPLES OF PHYLOGENETICS" Spring 2006
University of California, Berkeley B.D. Mishler

April 20, 2006. **Phylogenetic Trees IX: Summary -- what are they, really, and what can go wrong?**

A. Lead-off discussion

We have now reached the point in the course where we have covered both the character analysis phase and the cladistic analysis phase of phylogenetics in great detail. So let's step back and look at these phases together in a more integrated, sophisticated way. Here are some important initial questions for discussion:

What are phylogenetic trees, really?

What do you see when you look closely at a branch?
-- the fractal nature of phylogeny (is there a smallest level?)

What is the relationship between characters and trees? Characters and OTUs? Characters and levels?
-- exemplar coding versus composite coding

What can go wrong in the fit between characters and trees?
-- random versus caused homoplasy
-- epistemological problems versus ontological problems

B. Summary of character analysis

The process of phylogenetic analysis 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. 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. Extensive series of publications have been elaborated to attempt to justify such tree-building approaches as neighbor-joining, maximum likelihood, and Bayesian inference, while ignoring entirely the nature of the data matrix that must underlie any analysis. The reasons for this asymmetry in research on phylogenetic theory are not entirely clear, but it probably has to do with the fact that the problem of tree building may appear simpler, more clear cut. Perhaps it is just a matter of research fashions. For whatever reason, relatively little attention has been paid to the assembly of the data matrix. At stake are each of the logical elements of the data matrix: the *rows* (what are OTUs?), the *columns* (what are characters?), and the individual *entries* (what are character states?).

The tree of life is inherently fractal, which complicates the search for answers to these questions. Look closely at one lineage of a phylogeny and it dissolves into many separate lineages, and so on down to a very fine scale. Thus the nature of both OTU's ("operational taxonomic units," the "twigs" of the tree in any particular analysis) and characters (hypotheses of homology, markers that serve as evidence for the past existence of a lineage) change as one goes up and down this fractal scale. Furthermore, there is a tight interrelationship between OTUs and character states, since they are reciprocally recognized during the chacter analysis process.

C. Summary of tree building

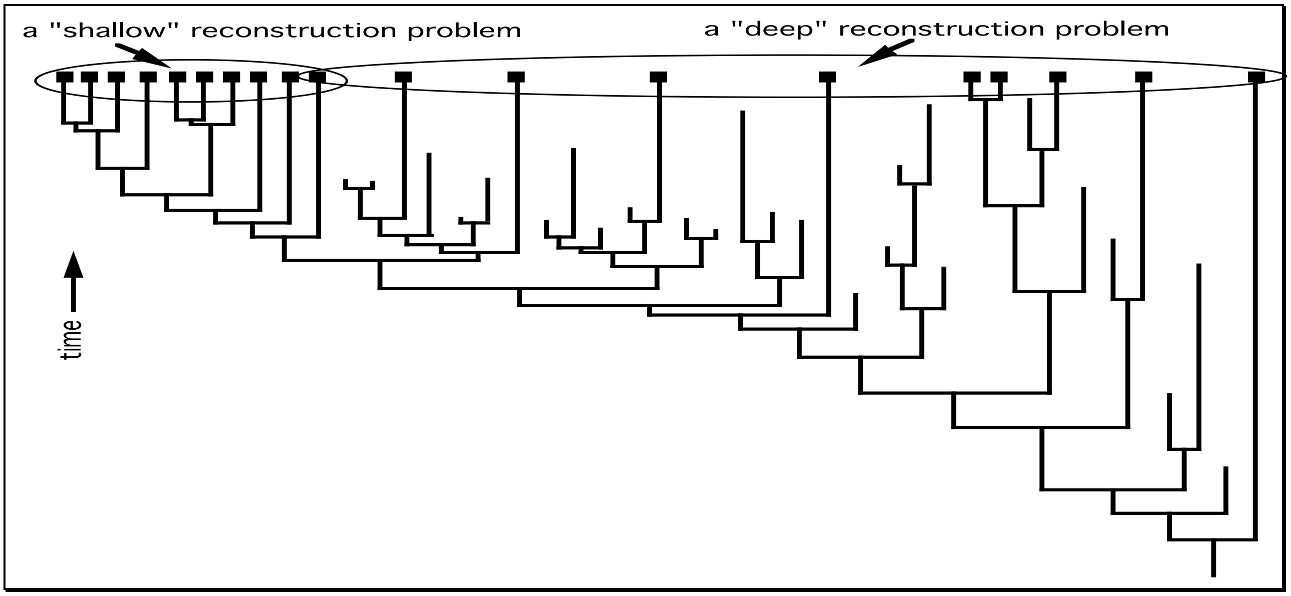
One could easily argue that the first phase of phylogenetic analysis is the most important; the tree is basically just a re-representation of the data matrix with no value added. This is especially true from a parsimony viewpoint, the point of which is to maintain an isomorphism between a data matrix and a cladogram. Under this viewpoint, we should be very cautious of any attempt to add something beyond the data in translating a matrix into a tree! If care is taken to construct an appropriate data matrix to address a particular question of relationships at a given level, then simple parsimony analysis is all that is needed to transform a matrix into a tree. Debates over more complicated models for tree-building can then be seen for what they are: attempts to compensate for marginal data.

But what if we need to push the envelope and use data that are questionably suited for a particular problem? More complicated model-based methods (weighted parsimony, ML, and Bayesian inference) can be used to push the utility of data, but need to be done carefully. Both the model itself and the values for the parameters in the model need to be based on solid a priori evidence, not inferred ad hoc solely from the data to be used.

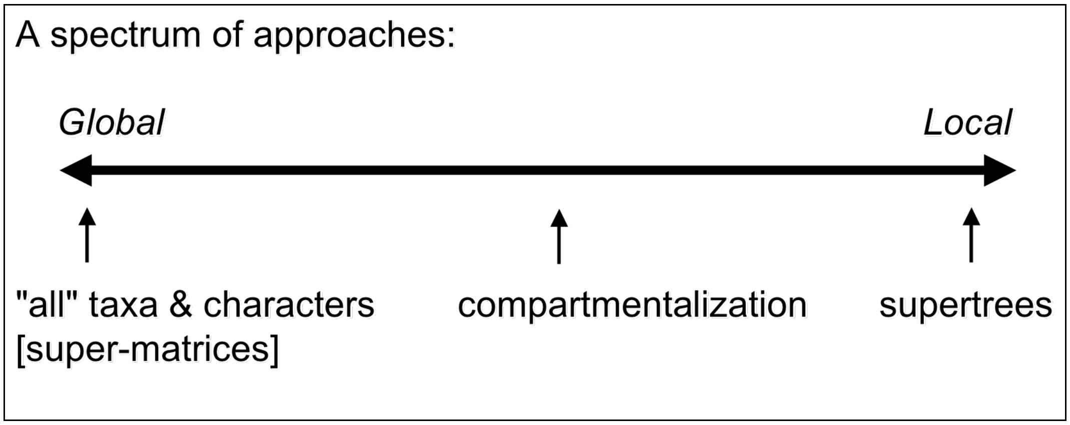
D. **"Deep" versus "shallow" phylogenetic inference: molecules and morphology**

The problems faced at different temporal scales are quite distinct (Mishler, 2000. *Taxon* 49: 661-683). In "shallow" reconstruction problems, the branching events at issue happened a relatively short time ago and the set of lineages resulting from these branching events is relatively complete (extinction has not had time to be a major effect). In these situations the relative lengths of internal and external branches are similar, giving less opportunity for long branch attraction. However, the investigator working at this level has to deal with the potential confounding effects of reticulation and lineage sorting. Characters, at least at the morphological level, may be quite subtle, and at the nucleotide level it is necessary to look very carefully to find rapidly evolving genes (however, such genes are likely to be relatively neutral, thus less subject to adaptive constraints which can lead to non independence).

In "deep" reconstruction problems, the branching events at issue happened a relatively long time ago and the set of lineages resulting from these branching events is relatively incomplete (extinction has had a major effect). In these situations, the relative lengths of internal and external branches are often quite different, thus there is more opportunity for long branch attraction, even though there is little to no problem with reticulation and lineage sorting since most of the remaining branches are so old and widely separated in time. Due to all the time available on many branches, many potential morphological characters should be available, yet they may have changed so much as to make homology assessments difficult; the same is true at the nucleotide level, where multiple mutations in the same region may make alignment difficult. Thus very slowly evolving genes must be found, but that very conservatism is caused by strong selective constraints which increases the danger of convergence.


a "shallow" reconstruction problem
a "deep" reconstruction problem
time

How will we ultimately connect up "deep" and "shallow" analyses, each with their own distinctively useful data and worrisome problems? Some hold out hope for eventual global analyses someday, once enough universally comparable data have been gained and computer programs get much more efficient, that can deal with all extant species at once, thus breaking down the conceptual difference presented above. Others would go to the opposite extreme, and use the "supertree" approach, where the "shallow" analyses are simply grafted onto the tips of the "deep" analyses (e.g., the "shallow" analysis in the figure above can be grafted onto the "deep" analysis there because of the single shared species between analyses). I favor an intermediate approach, called "compartmentalization" (Mishler, 1994; Mishler et al., 1998), where the "shallow" topologies (that are based on analyses of the characters useful locally) are imposed as constraints in global "deep" analyses (that are based on analyses of characters useful globally).



A spectrum of approaches:

Global *Local*

"all" taxa & characters [super-matrices]

compartmentalization

supertrees

E. Compartmentalization

This new and still controversial approach, called compartmentalization by analogy to a water-tight compartment on a ship (homoplasy is not allowed in or out), involves substituting an inferred "archetype" or hypothetical ancestor for a clade accepted as monophyletic *a priori* into an inclusive analysis (Mishler, 1994, *American Journal of Physical Anthropology* 94: 143-156).

It differs from the exemplar approach in that the representative character-states coded for the archetype are based on all the taxa in the compartment (thus the archetype is likely to be different from all the real taxa). In brief, the procedure is to: (1) perform global analyses, determine the best supported clades (these are the compartments); (2) perform local analyses within compartments, often with augmented data sets (since more characters can usually be used within compartments due to the improved homology assessments, as discussed below); (3) return to a global analyses, in one of two ways, either (a) with compartments represented by single OTUs (the archetypes), or (b) with compartments constrained to the topology found in local analyses (for smaller data sets -- this approach is better because it allows the character states of the archetypes to "float" with character optimizations based on the overall tree topology).

The goals of compartmentalization are to cut large data sets down to manageable size (the most obvious effect, but not the most important theoretically), suppress the effect of "spurious" homoplasy, and allow use of more information in analyses. The last is the most subtle point (but probably the most important) -- improved homology assessments can be made within compartments. This has been instinctively done by morphologists; when characters are being defined, only the "relevant" organisms (i.e., previously accepted as related) are compared (e.g., leaf-cell size is an important cladistic character within the moss genus Tortula, yet obviously this character would have to be eliminated if character-state divisions had to be justified across all the mosses together). There are also analogous advantages in molecular data. Alignments can be done more easily, and most accurately, when closely related organisms are compared first (Mindell, 1991) . Regions that are too variable to be used globally (and thus must be excluded from a global analysis) can often be aligned and included in a local analysis within a compartment. These goals are self-reinforcing; as better understanding of phylogeny is gained, the support for compartments will be improved, leading in turn to refined understanding of appropriate characters and OTUs.

F. Personal Summary

These issues of how to use phylogenetic markers at their appropriate level to reconstruct the extremely fractal tree of life are likely to be one of the major concerns of the theory of phylogenetics in coming years. In the future, my prediction is that more careful selection of characters for aparticular questions, that is more careful and rigorous construction of the data matrix, will lead to less emphasis on the need for modifications to equally-weighted parsimony. The future of phylogenetic analysis appears to be in careful selection of appropriate characters (discrete, heritable, independent, and with a low λ) for use at a carefully defined phylogenetic level. To paraphrase the New York Times masthead, we should include "all the characters that are fit to use."

What is the relationship between my emphasis on the data matrix, and general preference for parsimony? Simple: A rigorously produced data matrix has already been evaluated carefully for potential homology of each feature when being assembled. Everything interesting has already been encoded in the matrix; what is needed is a simple transformation of that matrix into a tree without any pretended "value added." Straight, evenly-weighted parsimony is to be preferred, because it is a robust method (insensitive to variation over a broad range of possible biasing factors) and because it is based on a simple, interpretable, and generally applicable model.

Data first!