

March 18, 2008. **Phylogenetic Trees I: Reconstruction; Models, Algorithms & Assumptions**

**A. Trees -- what are they, really, and what can go wrong?**

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?

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 character analysis process.

**B. Two approaches to tree-building**

What is the basic goal of tree building? How good is the fit between "reality" and a phylogenetic model designed to represent reality? These questions have many different answers depending on the background of the investigator, but there are two major schools of thought:

*1. The "reconstruction" school of thought.*

The Hennigian phylogenetic systematics tradition, derived from comparative anatomy and morphology, focuses on the implications of individual homologies. This tradition tends to conceive of the inference process as one of reconstructing history following deductive-analytic procedures. The goal is seen as coming up with the best supported hypothesis to explain a unique past event.

-- the data matrix as itself a refined result of character analysis

-- each character is an independent hypothesis of taxic and transformation homology

-- test these independent hypotheses against each other, look for the best-fitting joint hypothesis

-- straight parsimony as a "solution" to the data matrix

-- only the fewest and least controversial assumptions should be used: characters are heritable and independent, and that changes in state are relatively slow as compared to branching events in a lineage

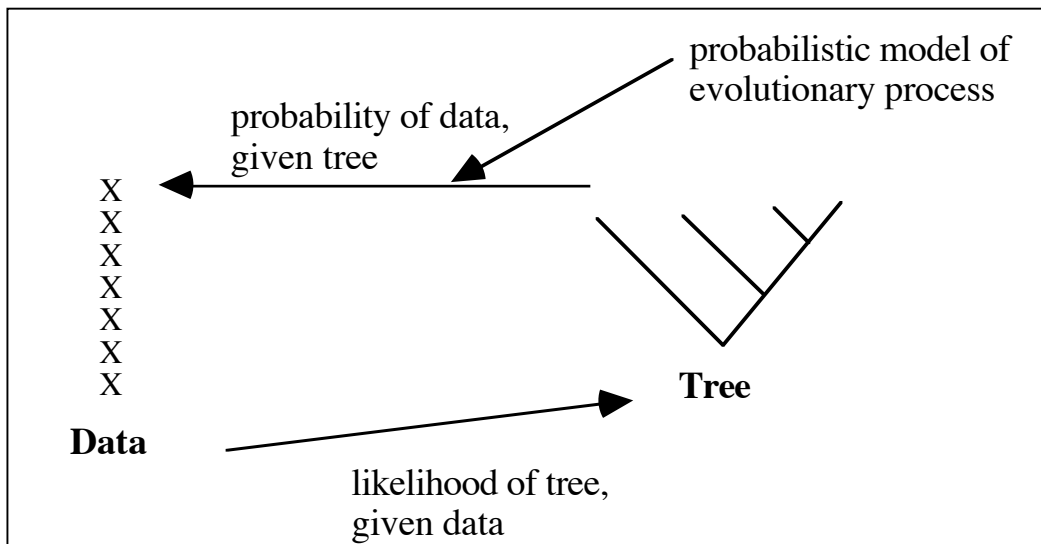
-- when these hold, reconstructions for a character showing one change on one branch will be more likely than reconstructions showing two or more changes in that character on different branches.

## 2. The "estimation" school of thought

The population genetic tradition, derived from studies of the fate of genes in populations, tends to see phylogenetic inference as a statistical estimation problem. The goal is seen to be choosing a set of trees out of a statistical universe of possible trees, while putting confidence limits on the choice.

-- task is to pick the single tree out of the statistical universe of possible trees that is the most *likely* given the data set.

--relationship between probability and likelihood (see figure below)



A maximum likelihood approach to phylogenetic estimation attempts to evaluate the probability of observing a particular set of data, given an underlying phylogenetic tree (assuming an evolutionary model). Among competing phylogenetic trees, the most believable (likeliest) tree is one that makes the observed data most probable.

-- to make such a connection between data and trees, it is necessary to have auxiliary assumptions about such parameters as the rate of character change, the length of branches, the number of possible character-states, and relative probabilities of change from one state to another. Hence, there is controversy.

The procedure (more details after break!)

- You need three things: Data, a Model, and a Likelihood Function.
- The Data is our normal matrix, where each column is a vector.
- The Model has three parts:
  1. a topology
  2. branch lengths (# of changes)
  3. model of changes (nucleotide substitution model, base frequencies, among-site variation)
- The Likelihood Function begins with the evaluation of each character, one at a time, considering the probabilities of all possible assignments of states to the internodes. The overall likelihood is the sum of the likelihoods of all the characters.

### **C. The role of statistics in phylogenetics?**

**\*\*There is a need to be clear about what statistical approaches are appropriate for a particular situation, or even whether any such approach is appropriate.\*\***

1. There are many schools of thought in statistics, but the general goal is a statement of uncertainty about hypotheses. The two schools of thought discussed above have different views about the role of stats, given their different approaches to epistemology.

2. The jury is still out on the applicability of various statistical approaches (or even the desirability of such approaches). Issues under debate include:

a. The nature of the statistical universe being sampled and exactly what evolutionary assumptions are safe to use in hypothesis testing. Under standard views of hypothesis testing, one is interested in evaluating an estimate of some real but unknown parameter, based on samples taken from a relevant class of individual objects (the statistical universe).

b. It might be argued that a particular phylogeny is one of many possible topologies, thus somehow one might talk about the probability of existence of that topology or of some particular branches. However, phylogenies are unique historical events ("individuals" in the sense of Hull, 1980) ; a particular phylogeny clearly is a member of a statistical universe of one. It is of course valid to try to set a frequency-based probability for such phylogenetic questions as: How often should we expect to find completely pectinate cladograms? or How often should we find a clade as well supported as the mammals? In such cases, there is a valid reference class ("natural kind" in the sense of Hull, 1980) about which one can attempt an inference.

c. It could be reasonably argued that characters in a particular group of organisms are sampled from a universe of possible characters. Widely-used data re-sampling methods pioneered by Felsenstein (jackknife and bootstrap) are based on this premise. The counter-argument, however, is that characters are chosen based on a refined set of criteria of likely informativeness, e.g., presence of discrete states, invariance within OTUs, ability to determine potential homology (including alignability for molecular data). Therefore, the characters are at best a highly non-random sample of the possible descriptors of the organisms. It may perhaps be better not to view characters as a sample from a larger universe at all -- a data matrix is (or at least should be) all the "good" characters available to the systematist.

d. Simulation approaches (i.e., building known trees using Monte Carlo methods and then generating a data set by evolution on that tree) are being used to understand how well different methods recover the truth under different circumstances, but they are of course very sensitive to our expectations about real phylogenies. What are proper null models for evolutionary tree? A difficult question to address because expected character distributions vary depending on tree topology and mode of character evolution. We can design a method to work well on a given known situation, but how do you know what method to pick for an actual study when you don't know what has happened in the past?

### 3. My own view:

a. Statistical considerations primarily enter systematics during the phase called "character analysis," that is when the data matrix is being assembled. Based on expectations of "good" phylogenetic markers (characters), procedures have been developed that involve assessing the likely independence and evolutionary conservatism of potential characters using experimental and statistical manipulations.

b. By the time a matrix is assembled, each column can be regarded as an independently justified hypothesis about phylogenetic grouping, an individual piece of evidence for the existence of a monophyletic group (a putative taxic homology). The parsimony method used to produce a cladogram from a matrix should then be viewed as a solution of that matrix, an analytic transformation of the information contained therein from one form to another, just as in the solution of a set of linear equations. No inductive, statistical inference has been made at that step, only a deductive, mathematical one. Now to assert that the resulting cladogram represents a model of a phylogenetic tree is another matter, an inductive inference requiring separate justification.

c. Maximum likelihood techniques remain the preferred statistical approach for such problems. A maximum likelihood approach attempts to evaluate the probability of observing a particular set of data, given an underlying phylogenetic tree. Among competing phylogenetic trees, the most believable (likeliest) tree is one that makes the observed data most probable. To make such a connection between data and trees, however, it is necessary to have auxiliary assumptions about such parameters as the rate of character change, the length of branches, the number of possible character-states, and relative probabilities of change from one state to another. The primary debate has involved these assumptions -- **how much is necessary or desirable or possible to assume about evolution before a phylogeny can be established?**

Sober (1988) has shown convincingly that some evolutionary assumptions are necessary to justify any method of inference, but he (and the field in general) remains unclear about exactly what the minimum assumptions are. Keep in mind also that Parsimony and Likelihood are fundamentally related methods -- a spectrum rather than two distinct methods. [More below and in future lectures]

d. It seems generally agreed that only the fewest and least controversial assumptions should be used. Given its assumptions as discussed above, the Wagner parsimony method appears to give a robust connection between data and preferred tree(s). In other words, assuming that characters are heritable and independent, and that changes in state are relatively slow as compared to branching events in a lineage, reconstructions for a character showing one change on one branch will be more likely than reconstructions showing two or more changes on different branches.

#### **D. When do straight parsimony methods fail?**

##### 1. The Felsenstein Zone (the central parameter $\lambda$ ):

The best way to predict phylogenetic behavior of characters (i.e., those that otherwise meet the criteria of detailed similarity, heritability, and independence) is by examining variation in the central parameter  $\lambda$ , defined as branch length in terms of expected number of character changes per branch [segment] of a tree. The advantage of using this parameter rather than the more commonly used "rate of character change per unit time" is that the former measure incorporates both rate of change per unit time and the length of time over which the branch existed. Thus, a high  $\lambda$  can be due to either a high rate of change or a historically long branch (both have an equivalent effect on parsimony reconstruction). This parameter, either for a single character, or averaged over a number of characters, defines a "window of informativeness" for that data. In other words, a very low value of  $\lambda$  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  $\lambda$  indicates data that are changing so frequently that problems arise with homoplasy through multiple changes in the same character. At best a high  $\lambda$  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.

The effects of differential  $\lambda$  values have been investigated by several workers. In an important early paper, 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 in parsimony reconstructions using a large number of characters (see also Sober, 1988). The region where branch-length asymmetries will tend to cause such problems has been called the "Felsenstein Zone". The seriousness of this problem (i.e., the size of the Felsenstein Zone) is affected by several factors, the most important of which are: (i) the number of possible character states per character; and (ii) the overall rate of change of characters.

## 2. How to "push back" the boundaries of the Felsenstein Zone?

- Selection and definition of OTUs and characters
- Additional taxa (which taxa?)
- Additional characters (which characters?)

## 3. Weighting issues:

Could there be feed-back from these considerations into methods for reconstructing phylogenies? As discussed above, maximum-likelihood considerations show that straight unweighted parsimony will provide a very close approximation (better at increasingly lower  $\lambda$ s), that may, however, require some adjustment when large asymmetries exist (and can be specified) in transformation probabilities among characters, among states within a character, or both. This adjustment to "straight parsimony" can be made via appropriate character and character-state weights. If differential  $\lambda$ 's for different characters (or types of characters) can be discovered a priori, then weights can be specified (e.g., weights taking into account differential probabilities of change at different codon positions in a protein-coding gene). Differential probabilities of transformation that can be specified among states within characters can be modeled similarly (e.g., weights taking into account gains versus losses in restriction site data, or transition/transversion bias in sequence data; see Albert, et al., 1993; Albert and Mishler, 1992; Albert, et al., 1992).

What if there are valid reasons for not viewing all apparent taxic homologies as equal in the weight of evidence they bring to the analysis? What, exactly, could those reasons be?

Possibilities for weighting include:

- (1) *A posteriori* weighting (e.g., Farris's successive approximations method)
- (2) *A priori* weighting (i.e., based on data external to those being used to infer a particular phylogeny); comes in two flavors:
  - i. character weights
  - ii. character-state weights

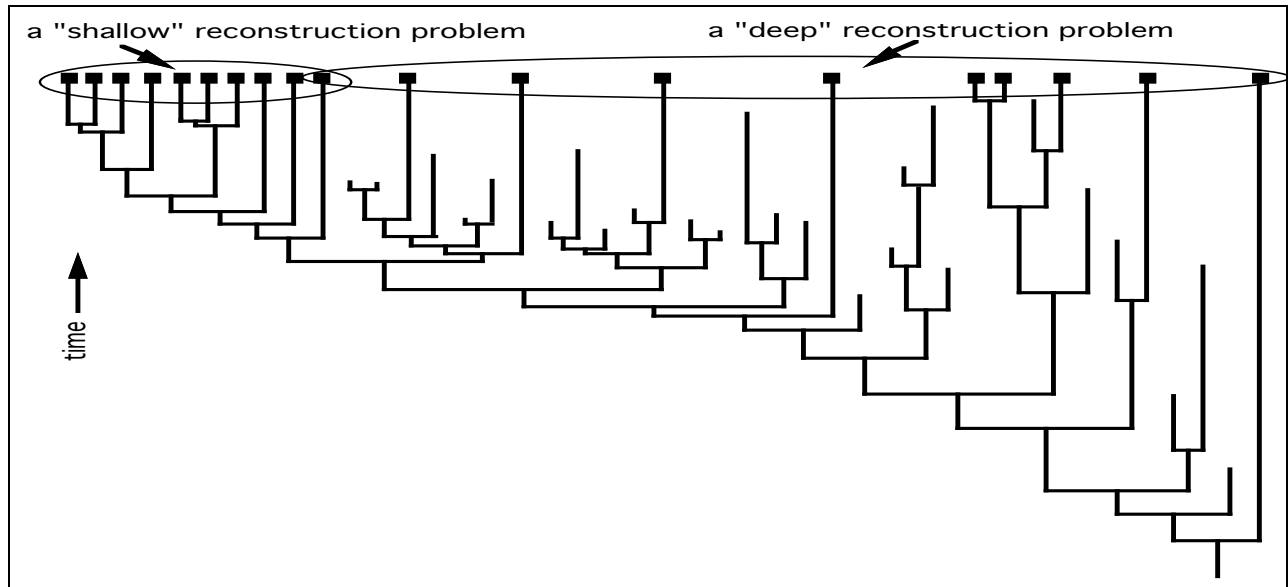
If differential  $\lambda$ 's for different characters (or types of characters) can be discovered a priori, then maximum likelihood-based weights can be specified (e.g., weights taking into account differential probabilities of change at different codon positions in a protein-coding gene). This is a simple matter of introducing a multiplier representing the relative weight. The relative weight of a character is the negative natural log of its relative probability of change (so high probability of change = low weight).

Specifying differential probabilities of transformation among states within characters is a little more difficult algorithmically, but can be done similarly (e.g., weights taking into account gains versus losses in restriction site data, or transition/transversion bias in sequence data). The method for applying such character-state weights is a step matrix. This specifies the "cost" of going from one state to another, and can be very complex (even asymmetrical).

It obviously is difficult to specify expectations for  $\lambda$  before an analysis; currently such approaches can only be attempted for molecular data (one advantage of its relative simplicity), therefore we are far from being able to use this sort of approach for combining molecular and morphological data. Fortunately, one important conclusion of our attempts at modeling the major known transformational asymmetries is that the differential weights thus produced have little effect on parsimony reconstructions. With data having a reasonable  $\lambda$  ( $\leq 0.1$ , as will be discussed in more detail in a later lecture), optimal weighted parsimony topologies are usually a subset of the unweighted (or more properly, equally-weighted) ones. Thus, paradoxically, our pursuit of well-supported weighting schemes ended up convincing us of the broad applicability and robustness of equally-weighted parsimony.

### E. "Deep" versus "shallow" phylogenetic inference: molecules and morphology

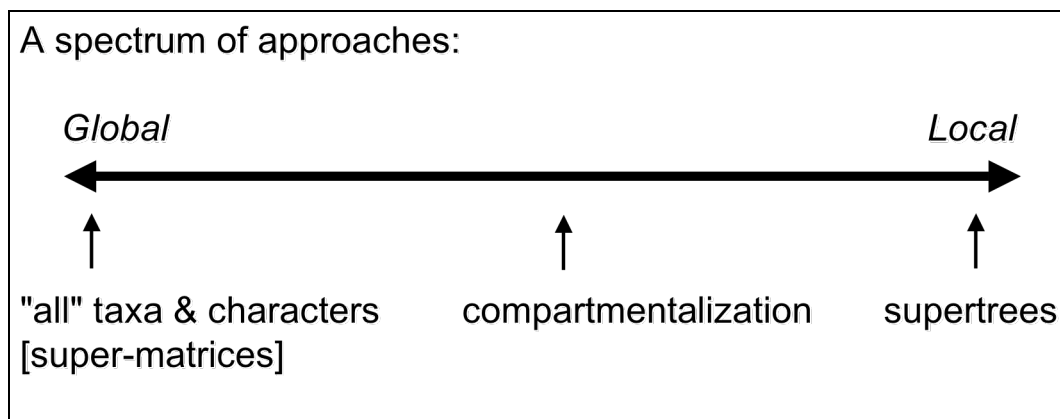
The problems faced at different temporal scales are quite distinct (Mishler, 2000. *Taxon* 49: 661-683). In "shallow" reconstruction problems, the branching events at issue happened a relatively short time ago and the set of lineages resulting from these branching events is relatively complete (extinction 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.

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).



## F. 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.

### **G. Summary:**

It is clear that parsimony works best with "good" data, i.e., with copious, independent, historically informative characters (homologies), evenly distributed across all the branches of the true phylogeny. Indeed, many competing methods tend to converge in their results with such data. It is in more problematic data (e.g., with limited information, a high rate of change, or strong functional constraints) that results of different methods begin to diverge. Data that are marginal or poor will be problematic for any approach, but different approaches account for (or are affected by) "noise" differently. Weighting algorithms may be able to extend the "window of informativeness" for problematic data, but only if the evolutionary parameters that are biasing rates of change are known.

One could easily argue that the character analysis 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.

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 a particular 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  $\lambda$ ) 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 my personal 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!*