Feb. 5, 2016. **Phylogenetic Trees I: Reconstruction; Models, Algorithms & Assumptions**
Reading assignment: *Tree Thinking* pp 35-53.

I. Summary of previous lectures on homology and characters, and goals of this one:

Hennigian phylogenetics can be most tersely described as the study of homology and its implications (Patterson, 1982). The basic criteria of character analysis, discussed the last few lectures, amount to a joint assumption that an apparent taxic homology [N.B., this a feature that has already passed strict observational and experimental tests of detailed similarity, heritability, discrete states, and independence] is more likely to be due to true taxic homology than to homoplasy, unless evidence to the contrary exists, i.e., a majority of apparent taxic homologies showing a different pattern. We assemble a matrix of hypothesized homologies, and evaluate them relative to each other. This requires that we can build well-supported phylogenetic trees from the matrix, the subject to introduce today.

II. 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.

III. Tree-building; Algorithms & Assumptions; reconstruction vs. estimation ??

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.

-- Statistical considerations primarily enter the process 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.

-- This school of thought tends not to see the tree building process per se to involve statistical inference. Since each column in the data matrix is 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 is then viewed as a <u>solution</u> 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.

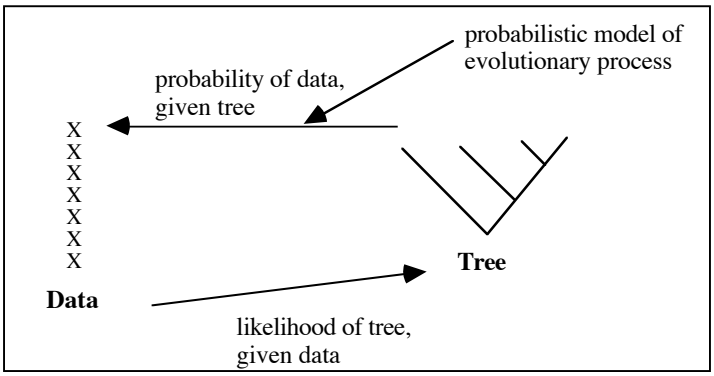
-- In summary: a rigorously produced data matrix has already been evaluated carefully for potential homology of each feature when being assembled. Everything interesting has already been encoded in the matrix; what is needed is a simple transformation of that matrix into a tree without any pretended "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.

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 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 <u>some</u> evolutionary assumptions are necessary to justify any method of inference, but he (and the field in general) remains unclear about exactly <u>what</u> the minimum assumptions are or should be. Keep in mind also that parsimony and likelihood are fundamentally related methods -- a spectrum of character-based methods rather than two distinct methods. [More in future lectures]

The procedure (more details in later lecture!)

-- 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.

IV. The role of statistics in phylogenetics?

****There is a need to think clearly 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) <u>all</u> 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

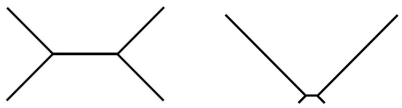
known situation, but how do you know what method to pick for an actual study when you don't what has happened in the past?

V. When do phylogenetic methods fail?

1. The Felsenstein Zone (the central parameter λ):

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 λ, defined as the expected number of character changes per branch [segment] of a tree (see: V.A. Albert, M.W. Chase, and B.D. Mishler. 1993. Character-state weighting for cladistic analysis of protein-coding DNA sequences. Annals Missouri Botanical Garden 80: 752-766). 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 λ 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 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.

The effects of differential λ 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.



"good" "bad"

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

A. Selection and definition of OTUs and characters

B. Additional taxa (which taxa?)

C. Additional characters (which characters?)

D. Modeling approaches:

-- incorporate more complicated models to take into account known biasing factors, e.g. differential probabilities of change at different codon positions in a protein-coding gene, transition/transversion bias, gains versus losses in restriction site data, or branch length asymmetry.

-- Where do models come from?

-- Where do values for parameters in models come from?

(1) from the data at hand? (*a posteriori*)

(2) from data external to those being used to infer a particular phylogeny? (*a priori*)

VII. Conclusions

It is clear that phylogenetic methods work best with "good" data, i.e., with copious, independent, historically informative characters (homologies), evenly distributed across all the branches of the true phylogeny, evolving at an appropriate rate for the depth of the problem. Most 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 in parsimony, or maximum likelihood methods 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. 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 phylogenetic 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 complicated modeling. The future of phylogenetic analysis appears to be in careful selection of appropriate characters (discrete, heritable, independent, and with the right rate of change for the time depth under study) for use at a carefully defined phylogenetic level.