"PRINCIPLES OF PHYLOGENETICS: ECOLOGY AND EVOLUTION"

Integrative Biology 200B Spring 2011

University of California, Berkeley B.D. Mishler

Jan. 20, 2011. **Phylogenetic reconstruction in a nutshell: trees**

I. Summary of previous lecture:

Hennigian phylogenetics can be most tersely described as the study of homology and its implications (Patterson, 1982). The basic criteria of character analysis, discussed last time, 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 for 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 best way to turn a matrix into a tree? This question has many different answers for different investigators, depending on their background.

A. Phenetic (distance)

- These methods work from an intermediate distance or similarity matrix

• Disadvantages:
  -- usually assumes molecular clock.
  -- many distance measures used are non-metric, therefore one can't interpret branch lengths in terms of actual evolutionary events (Euclidean distance vs. Manhattan Distance).
  -- hides homoplasy.
  -- throws away the information on individual characters that was so laboriously obtained.
• Advantages:
  -- ??? (at best able to mimic the results of a phylogenetic analysis)
  -- Avoid circularity by not considering evolution?
  -- Averaging across whole genome?
  -- Avoiding problem of reticulation? (some argue phenetic methods are OK below species level, as in the field of "phylogeography" -- more later in the class).
• Bottom line: distance methods not recommended for the purpose of hypothesizing phylogenies, although of course they are very useful for other tasks in ecology and evolution.

B. Cladistic or Phylogenetic (using the information from individual characters directly)

These methods build phylogenetic hypotheses directly from the data matrix described last time. There are two main schools of thought, that have converged from different historical beginnings:

• 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 "reconstruction" 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. *The "estimation" school of thought.*

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.

• Controversy remains on the applicability of various statistical approaches (or even the desirability of such approaches). Issues under debate include:

1. 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).

2. 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.

3. It could be reasonably argued that characters in a particular group of organisms are sampled from a universe of possible characters. 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.

IV. Parsimony

A. The "reconstruction" school of thought.
-- 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

B. Statistical considerations primarily enter parsimony analysis during "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 (as summarized last lecture).

C. Straight parsimony viewed as a "solution" to the data matrix

-- 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.

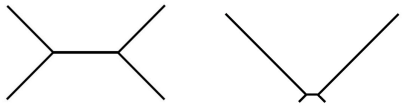
-- only the fewest and least controversial assumptions are used: characters are heritable and independent, and that changes in state are relatively slow as compared to branching events in a lineage

-- 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 (i.e., the Felsenstein Zone is avoided; see below).

D. 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 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 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 ($\bar{\lambda}$) defines a "window of informativeness" for that data. In other words, a very low value of $\bar{\lambda}$ 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 $\bar{\lambda}$ 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.



E. When do straight parsimony methods fail? 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?)
-- use more complex models for weighed parsimony, or maximum likelihood evaluations.

F. Character weighting:

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

Character and character-state weighting. Considerations of λ help to clarify character and character-state weighting (Albert, et al., 1993; Albert and Mishler, 1992; Albert, et al., 1992) . If differential λ '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).

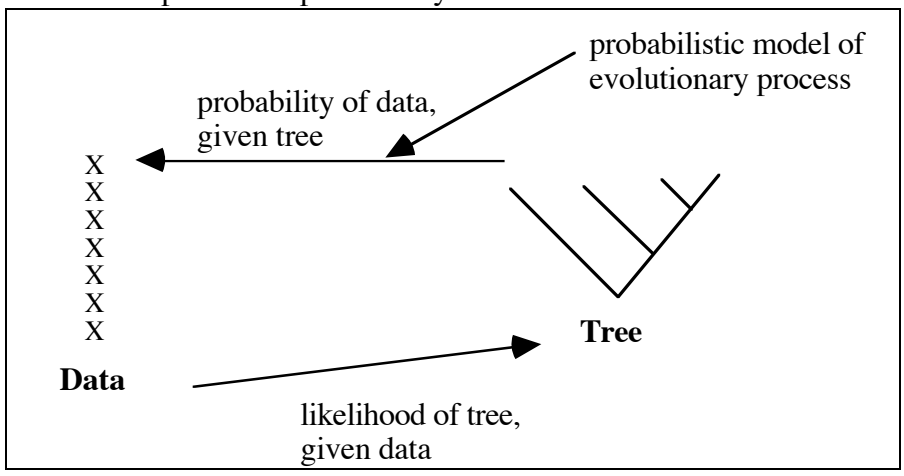
It obviously is difficult to specify expectations for λ before an analysis; currently such approaches can only be attempted for molecular data (one advantage of its relative simplicity), therefore we are far from being able to use this sort of approach for combining molecular and morphological data. 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 $\bar{\lambda}$ (≤ 0.1), 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 has ended up convincing us of the broad applicability and robustness of equally-weighted parsimony.

V. Maximum Likelihood

A. The "estimation" school of thought

-- 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 probablility and likelihood:



-- 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: **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 (and how many) assumptions are a good idea.

B. The procedure

-- You need three things: Data, a Model, and a Likelihood Function.

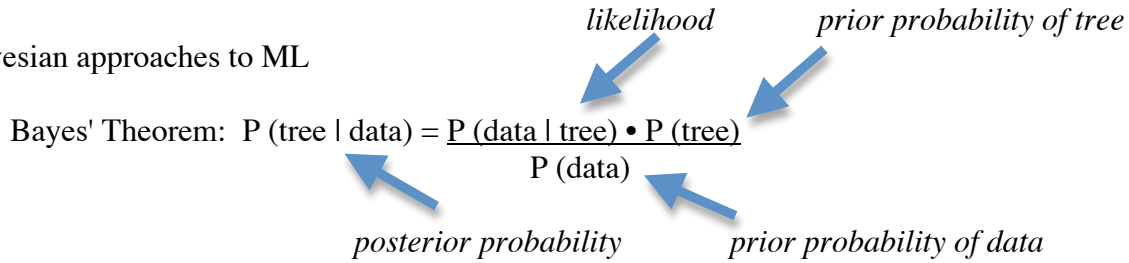
-- The Data is our normal matrix, where each column is a vector.

-- The Model has three parts:

1. a topology
2. branch lengths (# of changes)
3. model of changes (e.g. nucleotide substitution model, base frequencies, etc.)

-- The Likelihood Function begins with the evaluation of each character, one at a time, considering the probablilites of all possible assignments of states to the internodes. The overall likelihood is the sum of the likelihoods of all the characters. This is a big task, and algorithms tend to be very slow for decent-sized data sets.

C. Bayesian approaches to ML



-- Bayesian inference of phylogeny is based upon a quantity called the "posterior probability distribution of trees," which is the probability of trees conditioned on the observations using Bayes' Theorum as outlined above. The goal is to choose the set of trees having the highest posterior probability. If a particular monophyletic group appears in most or all of that set of trees, it deserves high credibility.

-- The posterior probability can't be calculated directly, so a heuristic approach is needed. Very efficient algorithms, using a Bayesian approach, have been developed to find regions of high likelihood (= high posterior probability) in tree space. The program MrBayes uses a simulation technique called Markov chain Monte Carlo (or MCMC) to approximate the posterior probabilities of trees (for more detail, see: http://mrbayes.csit.fsu.edu/).

D. Hypothesis testing.

-- the statistical nature of the ML estimation process lends itself to testing hypotheses about these variables (independent of controversies over its value as a reconstruction method). For example, the likelihood of different trees can be compared (using the likelihood ratio test statistic, LRT). We will thus revisit ML hypothesis testing later in the semester.

VI. Measures of Support

There are a number of connected topics that address the questions: How much faith should one put in one particular tree over another tree? How much faith should one put in one component of a tree over another component? More generally: How good is the fit between "reality" and the phylogenetic model designed to represent reality?

This is an active area of research; all that we can attempt here is an introduction to a few topics and a hint about the direction we think things should go.

A. Measures of fit.

1. One way to explore fit of data to trees is the *consistency index*, calculated as the minimum possible length of a tree divided by the actual length [**CI = M/S**].

-- Its advantage is the simple relationship to the parsimony criterion.

-- One problem with it as a measure of fit is that it is highly correlated with the number of taxa in a data set (Sanderson & Donoghue, 1989). This is related to problems discussed before: as taxa are added (with the number of possible states per character fixed), one would expect the number of random matches (= homoplasy) to increase. It is difficult to determine exactly what the minimum of the consistency index is, or what the CI for "random data" would be (due to problems with determining proper null hypotheses), but see Klassen et al. 1991 and Goloboff, 1991.

2. Another commonly used measure is the *retention index*, the fraction of apparent synapomorphy to actual synapomorphy [**RI = (G - S)/(G - M)** where G = min # of steps in the worst possible tree, a "star"]. The *rescaled consistency index* is the product of CI and RI

B. Robustness.

A way of assessing relative support of clades is also needed. Possibilities include:

(1) the number of characters at a node, or the number of "good" characters at a node (problem of optimization of characters -- ACCTRAN versus DELTRAN)

(2) bootstrapping (resampling replicate, full-sized data sets from the original)

(3) jackknifing (resampling replicate, smaller data sets from the original)

(4) likelihood (see above)

(5) posterior probability (see above)

(6) decay index (or Bremer Support). A non-statistical method, used with parsimony, based on the number of steps parsimony must be relaxed to make a particular clade lose its support. Using PAUP, can be calculated by obtaining the strict consensus of trees that are one step longer than the most parsimonious tree(s), then two steps longer, and so on until all

resolution is lost (use exhaustive search or "keep all trees ≤ length ___" option, then "filter trees"). Based on analysis of real and hypothetical data sets, this seems to be a sensitive measure of relative support.

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 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." In other words, DNA sequence data needs to be supplemented, or even replaced for some purposes, by morphological and structural-genomic data.