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

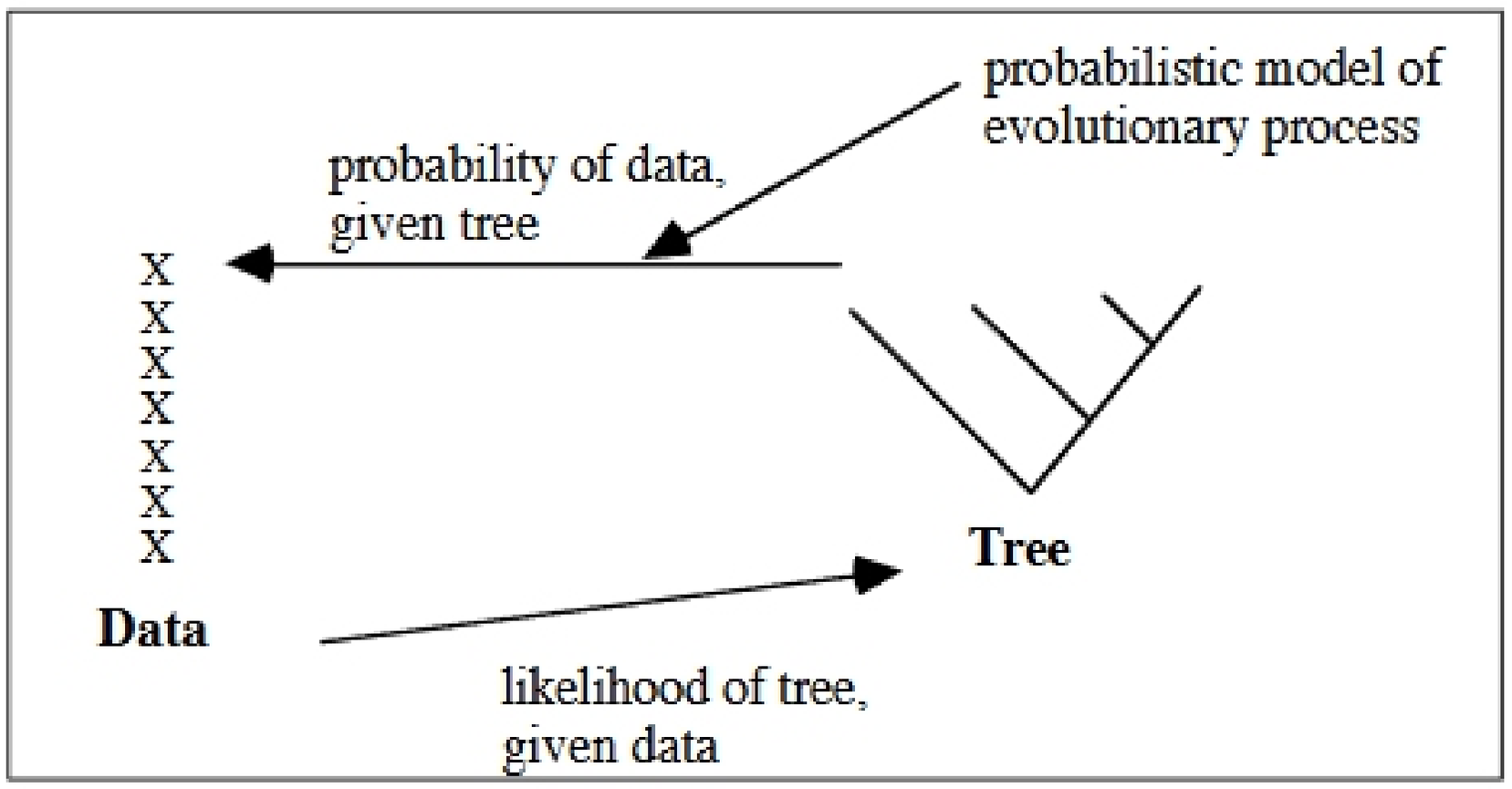
-- when these hold, reconstructions for a character showing one change on one branch will be more likely than reconstructions showing two or more changes in that character on different branches.

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 probablilites 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.