Feb. 12, 2020. **Phylogenetic trees III: Parsimony; Measures of support and robustness**

Reading assignment: *Tree Thinking* pp 95-106, 173-215, 271-284

1. Hennig and parsimony

Hennig was not concerned with parsimony as an optimality criterion, but rather his general paradigm was consistent with parsimony as a guiding principle (e.g. Occam's Razor as a heuristic rule of thumb). The connection is in Hennig's Auxiliary Principle – *to assume homology if there is no evidence to suggest otherwise*. Hennig provided fundamental methods that made the relationship between character evidence and cladograms explicit, but he did not provide a clear method for choosing among competing alternatives.

Parsimony as used in phylogenetics is often defined as "*minimizing evolutionary changes*." In one sense this is correct, but it should not be construed to mean that one thinks evolution proceeds by the fewest changes *possible* (but rather fewest changes *necessary*). If our character matrix consists of characters that have undergone rigorous character analysis to establish primary homology, we then should seek hypotheses (trees) that maximizing our homologies (and thus minimize homoplasy in those characters). We prefer trees that overturn as few as possible of our initial homologies, given that those initial hypotheses were the result of careful character analysis. The result is to *minimize ad hoc explanations* to account for failures in getting the primary homology hypothesis right.

There has been a tension among parsimony advocates between **pattern and process**, going back to the battles between numerical taxonomists and evolutionary systematists we talked about last time. Pattern cladists or transformed cladists are one extreme end of the spectrum. They put forward the idea that cladistic (in this case = strict parsimony) methods do not need, and in fact are better off without, an evolutionary (process) justification. According to them, only three things are needed to justify building trees based on synapomorphies: (1) discoverability of characters, (2) hierarchy as the best representation of the natural world, and (3) parsimony as an epistemological approach (Brower, A. 2000. Evolution is not a necessary assumption of cladistics. *Cladistics* 16, 143–154). Also part of the pattern v. process debates was the accusation of circularity, e.g. Mitter (1981. "Cladistics" in botany. Syst. Zool., 30:373–376.) "*there is widespread (but not universal) agreement that ... systematic methods should be as free as possible from assumptions about how evolution works, because these assumptions are in general not testable without reference to systematic results*." Much debate exists in the literature in regard to parsimony. Is it assumption free, assumption minimizing, or just a case where assumptions are ignored?

The clear understanding of patterns we observe and summarize, as opposed to processes that explain such patterns, is important (e.g. *multiple substitutions and insertions/deletions are inferred events, all processes, not observations*). However, a strict pattern view, which denies a role for evolution, does not provide a good explanation as to why any given character should or should not be included in an analysis or why we should use parsimony to infer trees.

For "reconstructionists" evolutionary models are definitely part of the character analysis and should be used (with caution) in tree building. In the "estimation school," maximum likelihood (ML), and the related Bayesian analysis (more later), rely heavily on evolutionary models for tree building. But exactly which model should be used, and especially where the values for parameters in the model should come from, remain major issues.

Parsimony and likelihood are best viewed as belonging to a family of methods. They are character based, using information about individual hypotheses of homology, unlike the distance methods we talked about last time. The connection between parsimony and likelihood is shown clearly in the case of the "no common mechanism" model (Penny et. al. 1994, Tuffley and Steel 1997). This model loosens the assumptions of rate change so that there is potentially a different rate for every combination of branch and character across the tree, which formally comes back to the parsimony model. Basically, parsimony has both the property of being the simplest model (all character and states treated equally) *and* the most complex model (each character assumed to have its own rate), thus the relationship between MP and ML is more like a circle than a spectrum. Another link between MP and ML is in character-weighted parsimony (below).

2. Parsimony as an optimality criterion

Goal: minimize the total number of steps over the tree. How do we measure steps (length)? A character has a *length* that is the number of independent changes in character states on a given cladogram. This is measured as steps or costs and may be differentially weighted (more below). There is frequently character conflict, i.e. character state distributions supporting groups that are not compatible. We hope to find the tree that has the shortest length summed across all characters. There are often multiple equally parsimonious trees that differ in topology. Also, even the same topology can have multiple, equally parsimonious inferences of character changes (alternative optimizations). ACCTRAN and DELTRAN are two "extremes" of optimization that may alter the implied transformational history of the character.

3. Weighting

Although the vast majority of people agree that all characters are not equally "good," equal-weights (which **is** a kind of weighting) is most commonly used.in parsimony, However, differential character weighting has always been part of parsimony. A weight is a factor applied to changes in characters to affect their length. There are two kinds of weighting: (1) character weighting (which applies a weight to all changes in some character), or (2) character-state weighting (which applies differential weights to changes in one character). The first acts the same as having multiple characters with the same state distribution in the matrix. The second uses the step matrix we talked about earlier to weight different transitions differently within a character.

As discussed in an earlier lecture, straight unweighted parsimony can provide an accurate reconstruction at lower rates of change (measured using λ, the expected number of character changes per branch). However, when approaching the boundary of the Felsenstein zone adjustment are required when large asymmetries exist (and can be specified) in transformation probabilities among characters, among states within a character, or both. This adjustment to straight, equally-weighted parsimony can be made via appropriate character and character-state weights. If differential λ'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). 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). References for further reading on these methods as applied in weighted parsimony:

V.A. Albert, B.D. Mishler, and M.W. Chase. 1992. Character-state weighting for restriction site data in phylogenetic reconstruction, with an example from chloroplast DNA. In P. Soltis, D. Soltis, and J. Doyle (eds.), *Molecular Systematics of Plants*, pp. 369-403. Chapman and Hall.

V.A. Albert and B.D. Mishler. 1992. On the rationale and utility of weighting nucleotide sequence data. *Cladistics* 8: 73-83.
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.

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). 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). Step matrices can be used for any number of transformation or weighting schemes

The character state **step-matrix**, or **cost matrix** (Sankoff, 1975) are illustrated below.

Unordered

From: \ To:	0	1	2	3
0	0	1	1	1
1	1	0	1	1
2	1	1	0	1
3	1	1	1	0

Ordered

From: \ To:	0	1	2	3
0	0	1	2	3
1	1	0	1	2
2	2	1	0	1
3	3	2	1	0

Irreversible

From: \ To:	0	1	2	3
0	0	1	2	3
1	∞	0	1	2
2	∞	∞	0	1
3	∞	∞	∞	0

Once the weighting model has been decided on, we still have to come up with the values for the parameters in the model. Where do these values come from? Two possibilities: (1) *A priori* evidence (i.e., based on data external to those being used to infer a particular phylogeny) or *a posteriori* evidence (i.e., patterns seen in the data set at hand, perhaps inferred during tree building, e.g., Farris's successive approximations method).

4. Tree space and tree searching

There are a huge number of possible cladograms for any modest number of OTUs, and finding them is a proven NP complete problem. Think of these in tree space, which is all the possible topologies with similar ones closer together.

A. Strategies typically used to find most parsimonious trees (MPTs) include *enumeration*, i.e., look at every possible cladogram, sum length of all characters on each, and pick the shortest one. Of course, this is not practical for any significant matrix. *Thus we need heuristic methods and methods to escape local optima…*

B. Basic strategy of a heuristic search:

a. Get a starting tree. This could be purely random or you may try for a "pretty good" tree to start. This can be done by randomly putting three OTUs in a network and adding a forth to the edge that creates the shortest four OTUs network and so on until all OUTs are joined. This does not guarantee a shortest tree, but it usually is not too bad for a start.

b. Take starting tree and make small rearrangements ("branch swapping") to get nearby trees (think of trees as being in tree-space with similar ones close together)

c. If one of these nearby trees is better (shorter) then retain it, discard the old one and make rearrangements on the new tree. Repeat till you can't find a shorter one.

d. This is a basic "greedy algorithm" that will lead you to a local optimum for sure, but it may not be the global optimal solution.

C. Some more advanced ways to search:

a. Character reweighting methods such as the "Parsimony Ratchet", where you randomly select 5-25% of the characters in the matrix and temporarily increase their weight, do branch

swapping using the reweighted matrix, then reset weights and calculate lengths on the set of trees that were found, keep the best trees and repeat. This lets you jump to very different parts of tree space and get out of local optima.

b. Simulated annealing is a method for wandering tree space using a Metropolis Algorithm that has a "temperature" parameter which dictates how severe the wandering permutations are and this decreases overtime and hopefully can wander to the global optima. Similar algorithms are used in Bayesian searches.

c. Tree Drifting is similar to annealing but uses a Relative Fit Functions (Goloboff as applied in the program TNT).

5. Measures of support and robustness under parsimony

One way to measure 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. 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

A way of comparing clades within a tree is also needed ("robustness"). 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 again)

(2) bootstrapping or jackknifing. Both make replicate datasets from the original one, the former by sampling with replacement, the latter by deleting a certain number of characters. A "pseudo" statistical method.

(3) Decay index or Bremer support. An explicitly non-statistical method, 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.

6. Background on meaning of the term "cladistics"

"parsimony" ≠ "cladistics"

Some people mistakenly equate "cladistics" with "parsimony," but Hennigian cladistics is primarily a philosophy of biological classification, emphasizing phylogeny and based on Hennig's logical approach to the evidence given by hypotheses of character evolution — recognizing monophyletic groups using shared, derived characters (homologies). Willi Hennig himself was not quantitative, so did not advocate any numerical analytical method.

As quantitative approaches were developed in cladistics, parsimony, weighted parsimony, and maximum likelihood methods were all explored early. Now we realize that these methods, together with Bayesian phylogenetic inference, make up a related family of homology and character-based methods — they are all cladistic methods.

The highly ranked journal of the Willi Hennig Society, *Cladistics*, covers all methods of phylogenetic analysis and many topics in evolution, ecology, biogeography, and systematics. "Cladistics" is better equated with "phylogenetics."