Parsimony & Likelihood [draft]

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 a sense this is correct, but it should not be construed to mean that one thinks evolution is parsimonious. If our character matrix consists of characters that have undergone rigorous character analysis to establish conjectural or primary homology, we then should seek hypotheses (trees) that maximizing our homologies. Conversely, we prefer trees that overturn as few as possible of our initial homologies, given that these initial hypotheses are the result of careful character analysis. The result is to *minimize ad hoc explanations* when we fail to get the primary homology right.

The two views that parsimony is "minimizing evolutionary changes" or "minimizing ad hoc explanations" is part of a larger tension between **pattern and process**.

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. Three things are needed to justify building trees based on synapomorphies, 1. discoverability of characters, 2. hierarchy is 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.*"

The clear understanding of patterns we observe and summarize, as opposed to processes that explain such patterns is important. 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. Much debate exists in the literature in regard to parsimony. Is it assumption free, assumption minimizing or just a case where assumptions are ignored?

Parsimony and likelihood are best viewed as belonging to a family of methods. They are character-matrix based, using information about individual hypotheses of homology, unlike the distance methods we talked about. 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 comes back to the

parsimony model. Essentially parsimony has both the property of being the simplest model (straightforward summing changes of observed states) and the most complex model.

2. Parsimony as an optimality criterion.

Unique to parsimony:

a. Shared character states are always evidence of shared ancestry unless the weight of evidence suggests otherwise. - Weights can be used to assign confidence in characters.
b. There is no reason to postulate unobserved changes of an apparently fixed character state except based on tree topology.
c. Trees with the fewest number of independent origins of shared character states are preferred. -We know that there is character conflict, i.e. character state distributions support groups that are not compatible.

Shared with other methods:

a. There is a single correct tree.
b. Character states originate and become fixed over time, marking history.
c. Characters are independent.
d. Analyses are done with characters and resultant trees have characters placed on them.

3. Fit characters to trees

How do we measure steps (length)? ---- A character has a length that is the number of independent origins of character states on any given cladogram. This is measured as steps or costs and is weighted depending on the model assumptions.

Two kinds of equally parsimonious trees: 1. same topology but different character state distributions (optimization) 2. Different topologies.

Optimizations: ACCTRAN and DELTRAN. Two "extremes" of optimization that may alter the resolution or the tree and the implied transformational history of the character.

4. Tree space and tree Searching

-We know that there are a huge number of possible cladograms for any modest number of OTUs.

Number of OTUs-	Number of rooted, resolved trees-
2	1
3	3
4	15
5	105
6	945
7	10395
10	34459455
20	$\sim 10^{21}$

This is a proven NP complete problem.

Strategies typically used to find most parsimonious trees (MPTs):

a) Enumeration- look at every possible topology, sum length of all characters keep the shortest tree(s).
b) Branch and Bound- initiates looking at every topology (branch) but discontinues at any as soon as they reach the point that they exceed the possible shortest length (bound). This will find the shortest trees, but is still slow.
c) Heuristics. Permutes characters and trees for search strategies that cover tree space and ovoid being

trapped in a local optimum.

5. Maximum Likelihood. Maximum likelihood is a general statistical method for estimating unknown parameters of a probability model. A parameter is some descriptor of the model. A familiar model might be the normal distribution of a population with two parameters: the mean and variance. In phylogenetics there are many parameters, including rates, differential transformation costs, and, most important, the tree itself.

Likelihood is defined to be a quantity proportional to the *probability of observing the data given the model*, P(D|M). Thus, if we have a model (i.e. the tree and parameters), we can calculate the probability the observations would have actually been observed as a function of the model. We then examine this likelihood function to see where it is greatest, and the value of the parameter of interests (usually the tree and/or branch lengths) at that point is the maximum likelihood estimate of the parameter.

6. Maximum likelihood as an optimality criterion.

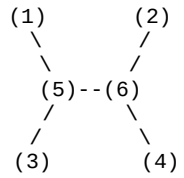
a) Appropriate for simple data like DNA sequences, where we can reasonably model the largely stochastic processes, i.e., a statistical description of the stochastic processes. But questionably applicable to complex data like morphology given the difficulty of modeling the numerous processes.
b) Better accounting for branch lengths, e.g. incorporates models of "multiple hits" thereby providing more realistic branch lengths. But the result is dependent on the model used and information derived from sites that are uninformative under parsimony is only due to the model used.
c) Lower variance than other methods (i.e., estimation method least affected by sampling error) and robust to many violations of the assumptions in the evolutionary model, even with very short sequences it may outperform alternative methods such as parsimony or distance methods. But may be susceptible to asymmetrical presence/absence of data in partitions *(see Simmons, M.P., 2011. Misleading results of likelihood-based phylogenetic analyses in the presence of missing data. Cladistics. 27:1-15.)*
d) the method is statistically well understood and evaluates different tree topologies and use all the sequence information but is very computationally intensive (though this is becoming much less of an issue)
e) Philosophically, especially in terms the applicability of probabilities and statistical measures of unique historical events (vs. Parsimony as a general principle). This is a fundamental distinction between reconstruction and estimation, e.g. *"Although the true phylogeny maybe "unknowable" it can nonetheless be estimated…"* Phylogenetic Inference", Swofford, Olsen, Waddell, and Hillis, in Molecular Systematics, 2nd ed., Sinauer Ass., Inc., 1996, Ch. 11.
f)

Simple tree example:

Assume that we have the aligned nucleotide sequences for four taxa:

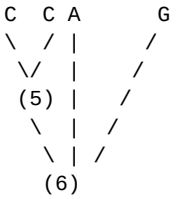
```
          1         j          N
(1)       A G G C T C C A A ....A
(2)       A G G T T C G A A ....A
(3)       A G C C C A G A A ....A
(4)       A T T T C G G A A ....C
```

and we want to evaluate the likelihood of the unrooted tree represented by the nucleotides of site **j**



What is the probability that this tree would have generated the data presented in the sequence under the chosen model?

Since most of the models currently used are **time-reversible**, the likelihood of the tree is generally independent of the position of the root. Therefore it is convenient to root the tree at an arbitrary internal node as done below.



Under the assumption that nucleotide sites evolve independently (the Markovian model), we can calculate the likelihood for each site separately and combine the likelihood into a total value at the end. To calculate the likelihood for site **j**, we have to consider all the possible scenarios by which the nucleotides present at the tips of the tree could have evolved. So the likelihood for a particular site is the summation of the probabilities of every possible reconstruction of ancestral states, given some model of base substitution. So in this specific case all possible nucleotides A, G, C, and T occupying nodes (5) and (6), or 4 x 4 = 16 possibilities.

Since any one of these scenarios could have led to the nucleotide configuration at the tip of the tree, we must calculate the probability of each and sum them to obtain the total probability for each site **j**.

The likelihood for the full tree then is the product of the likelihood at each site.

$$L = L(1) \times L(2) \ldots\ldots \times L(N) = \prod_{j=1}^{N} L(j)$$

Since the individual likelihoods are extremely small numbers it is convenient to sum the log likelihoods at each site and report the likelihood of the entire tree as the log likelihood.

$$\ln L = \ln L(1) + \ln L(2) \ldots\ldots + \ln L(N) = \sum_{j=1}^{N} \ln L(j)$$

Models:

Where μ= mean instantaneous substitution rate, when time is measured as substitutions μ=1

a,b,c,…l = rate parameter for each possible transformation of one base to another π_A= frequency of bases A, C,G, & T

transitions in bold

Nearly all substitution models are rate tables that are variation of this general form of the matrix below, e.g., JC sets all rates equal to 1 (a…l =1) and frequency are equal (π_A, π_C, π_G, π_T all equal ¼), K2P where the observation that transitions and transversions occur at different rates (b, e, h, k are adjusted by constant K).

The basic form is a matrix: Q =

	A	C	G	T
A	$-\mu(a\pi_C+b\pi_G+c\pi_T)$	$\mu a\pi_C$	$\boldsymbol{\mu b\pi_G}$	$\mu c\pi_T$
C	$\mu g\pi_A$	$-\mu(g\pi_A+d\pi_G+e\pi_T)$	$\mu d\pi_G$	$\boldsymbol{\mu e\pi_T}$
G	$\boldsymbol{\mu h\pi_A}$	$\mu i\pi_C$	$-\mu(h\pi_C+j\pi_G+f\pi_T)$	$\mu f\pi_T$
T	$\mu i\pi_A$	$\boldsymbol{\mu k\pi_C}$	$\mu l\pi_G$	$-\mu(i\pi_C+k\pi_G+l\pi_T)$

Among-Site Rate Variation (Γ)
The starting hypothesis is that all sites are assumed to have equal rates of substitution. This assumption can be relaxed, allowing rates to differ across sites by having rates drawn from a gamma distribution. The gamma is useful as its shape parameter (ά) has a strong influence on the values in the distribution.

Choosing a model:
As you might imagine, there are many models already available (ModelTest discussed below looks at >50!!) and an effectively infinite number are possible. How can one choose?

The program ModelTest (Posada & Crandal 1998) uses log likelihood scores to establish the model that best fits the data. Goodness of fit is tested using the likelihood ratio score.

$$\frac{\max [L_0 \text{ (simpler model)} \mid \text{Data}]}{\max [L_1 \text{ (more complex model)} \mid \text{Data}]}$$

This is a nested comparison (i.e. L_0 is a special case of L_1)

Adding additional parameters will always result in a higher likelihood score. However, at some point adding additional parameters is no longer justified in terms of significant improvement in fit of a model to a particular dataset. Over parameterizing simply fits the model to noise in the data.

A simple example:

HKY85 -lnL = 1787.08
GTR -lnL = 1784.82

Then, LR= 2*(lnL1-lnL2) ; LR = 2 (1787.08 - 1784.82) = 4.52

degrees of freedom = 4 (GTR adds 4 additional parameters to HKY85)
critical value (P = 0.05) = 9.49 -The added parameters are not justified by this significance test.