Introduction to Phylogeny, Trees and Parsimony as an Optimality Criterion

I. Roles of Phylogeny Estimation – Inference of the hierarchy of common descent.

An interest in understanding phlyogenetic relationships is as old as evolutionary biology.

The classical early examples of trying to diagram the tree of life date back to the well-known hypotheses of relationships presented by Heackel and Darwin.


How we currently estimate phylogenies is primary focus of this the first half of the course I’ll teach this spring with Luke, and lots of methods have been devised developed and tested.

It’s incredibly important – as of 10/01/08, a search of the ISI Web of Science database using “phylogen*” returned 62,801 papers published in this decade.

Examples of Uses (in addition to traditional use in classification):

Understanding Evolutionary Novelty.
Study of Evolutionary Patterns and Processes
  Co-speciation – for example between parasites and their hosts.
  Historical Biogeography – molecular phylogeography.
  Hybridization
  Gene Duplication and Genome Structure
  Macroevolutionary Patterns (Focus of 2nd half of course).
    extinction and speciation rates (e.g., lineage through time)
    species selection and the study of evolutionary trends
    key adaptations that allow a group to be speciose (ecological release).
    study of correlation of traits or characters – comparative biology

Conservation Biology
  Given the ongoing biodiversity crisis and the limited resources that society allocates to conservation issues, a need arises to prioritize conservation efforts. We saw a little of this in Laura Kubatko’s seminar yesterday, where phylogenies are being used in assessing taxonomy. In addition, phylogeny has been proposed as a means for determining these priorities through the concept of Phylogenetic Diversity (e.g., Faith, 1992. Biol. Cons. 61:1).


Forensic Evidence – Statistical phylogenetics have been used recently to connect a Louisiana physician to a patient who was the source of HIV used to infect his mistress intentionally (Metzker et al. 2002. PNAS, 99:14293). This is the first times that phylogenetics have been accepted as forensic evidence.
These are just a few of the many and varied uses of phylogenies, and you each could probably add to this list. The broad range of disciplines that this represents has a number of effects.

First, it leads to the diversity of backgrounds of scientists who use phylogenetics in their research – most are not phylogeneticists. If you look at papers that present some type of phylogeny, many of the analyses are very poorly done and sometimes results are simply misinterpreted.

Second, there has been a huge infusion of computer scientists and mathematicians into development of phylogenetic methods. We saw an excellent example of such a person yesterday during Laura Kubatko’s seminar, but sometimes these researchers are unaware of the historical development of the field, which has been contentious and absurdly vitriolic.

II. Some Preliminaries about Trees.

To begin our treatment of estimating phylogeny, I want to first introduce some generalities about trees, so that we’ll have a common starting point, and then discuss the various optimality criteria that one may choose for evaluating trees.

A. Some Basic Terms. There are two sets of terms that are commonly used: one by mathematicians and one by systematists.

In a sense, the mathematicians’ terms are more appropriate because phylogenetic trees fall into graph theory, however, having been trained as a systematist, I think in those terms.

1. Branches: In there are n-taxa in an unrooted tree, there are 2n-3 branches, labeled 1 – 7.
   - 1, 2, 4, 6, & 7 are external, or terminal branches (leaves in math jargon).
   - 3 & 5 are internal branches (edges).

2. Nodes are points where branches intersect or terminate

   Nodes A – E are terminal nodes, or terminals, whereas nodes x, y, & z are internal nodes.

   Nodes represent divergence events (speciation, gene duplications, coalescence, etc.).

3. We can denote trees in parenthetical notation (sometimes called New Hampshire trees).
   This would be represented by ((A,B),C,(D,E)).
4. We can break trees into subtrees. If we break branch 3, we have two subtrees (A,B) and (C,(D,E)).

B. Rooting - This is an unrooted tree.

This implies a lack of directionality, such that one may traverse the tree in any direction one finds convenient and start the traversal at any point.

Modern phylogenetic analyses usually estimate unrooted trees, and the root is usually determined by the inclusion of an outgroup. The point at which the outgroup connects with the ingroup is the root.

This is usually included in the estimation, not simply added after an ingroup topology is estimated (as I’ll do in this demonstration).

All of the rooted trees below can be derived from the unrooted tree by placing the root (common ancestor) on each branch. These rooted trees do imply directionality; nodes deeper in the tree represent divergence events that happened earlier. This illustrates how rapidly the number of alternative grows as one adds taxa. Addition of a single node, the root node (or alternatively, the addition of a single taxon, the outgroup), resulted in the seven trees below being derivable from the single tree above.
C. The Scope of the Phylogeny Problem.

One other important point to make right off at this early stage can be seen by the rooting exercise.

The number of possible trees increases incredibly fast as the number of taxa increases.

We can think of rooting as adding another taxon to this tree. Well, there are 7 places \((2n-3)\) we can add the 6th taxon and still be consistent with this unrooted tree.

What this means is the number of possible alternative trees will quickly be far too vast for us to examine every possible tree for a phylogenetic question involving very many taxa.

Number of possible trees for N taxa:

\[
B(N) = \prod_{i=3}^{N} (2i - 5)
\]

<table>
<thead>
<tr>
<th>Number of Taxa (i)</th>
<th>Number of Unrooted Trees</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>105</td>
</tr>
<tr>
<td>7</td>
<td>945</td>
</tr>
<tr>
<td>8</td>
<td>10,395</td>
</tr>
<tr>
<td>9</td>
<td>135,135</td>
</tr>
<tr>
<td>10</td>
<td>(2.027 \times 10^6)</td>
</tr>
<tr>
<td>22</td>
<td>(3 \times 10^{23})</td>
</tr>
<tr>
<td>50</td>
<td>(3 \times 10^{74})</td>
</tr>
<tr>
<td>100</td>
<td>(2 \times 10^{82})</td>
</tr>
<tr>
<td>1000</td>
<td>(2 \times 10^{2,860})</td>
</tr>
<tr>
<td>10,000,000</td>
<td>(5 \times 10^{68,667,340})</td>
</tr>
</tbody>
</table>

So the task of phylogeny inference is quite formidable, which is why the idea that phylogenetic trees can be made easy sticks in my craw.

Some approaches to phylogeny estimation don’t try at all to search among these alternatives, but rely on a series of rules to build one. These are referred to as algorithmic methods (although computer scientists hate this terminology).
Other approaches actually try to search this vast tree space in an effort to find the best tree under some optimality criterion.

### III. Optimality Criteria

Before we deal with how to choose from among all these trees, we first need to know how to evaluate any one tree.

Both algorithmic and optimality approaches rely on being able to evaluate a particular tree.

**A. There are three main criteria used for evaluating a tree.**

1. **Parsimony** – Given a character X taxon matrix, the tree that minimizes the number of character-state changes (across all characters) is the preferred tree.

2. **Minimum Evolution** – Given a matrix of pair-wise distances, the tree with the smallest sum of branch lengths is the preferred tree.

3. **Maximum Likelihood** – Given a dataset (an alignment) and a model of character evolution, the tree that has the highest probability of having generated the observed data is the preferred tree.

**B. Parsimony** – This is the most frequently used criterion. It’s certainly the easiest to compute, it’s also the only philosophically acceptable criterion for some systematists.

First, the parsimony score of a tree (i.e., its length) for the entire data set is given by:

\[ L_T = \sum_{i=1}^{s} \sum_{i=1}^{w_i l_i} \]

That is it’s the weighted sum of the character lengths, across all s characters. For our purposes, we’ll consider equal weights.

There are two ways one may calculate the \( l_i \)’s, I’ll just go over one here.

**The Fitch Algorithm** applies to matrix with all character-state transformations equal and operates on state sets and accumulated lengths.

---

![Diagram](image-url)
There are four terminals for this (arbitrarily rooted) tree: one with G, one with A and two with C. We erect a state set at each terminal node and assign an accumulated length of zero to terminal nodes. This is the minimum number of changes in the daughter subtree.

We proceed down the tree to internal nodes and calculate the state sets for each, as follows:

1 – Form the intersection of the two daughter nodes. If the intersection is non-empty, assign the intersection as the set for the internal node equal. The accumulated length of the internal node is the sum of those of the daughter nodes.

2 - If the intersection is empty, we assign the union of the two daughter nodes to the state set for the internal node. The accumulated length is the sum of those of the daughter nodes plus one.

Therefore, this character has a minimum length of 2 steps for this tree. In addition to determining the length of this tree for this character, we’ve also assigned possible ancestral states at each of the nodes.

Remember that this is only applicable to characters for which transformations among all characters are unordered and have the same cost.

C. Minimum Evolution

If the data are either inherently distance-based, or are converted to a series of pair-wise distances, one may choose to evaluate trees under the Minimum-Evolution criterion.

This matrix has \((n^2-n)/2\) cells (pair-wise distances).

Consider the following unrooted 4-taxon tree - There would be \((16-4)/2 = 6\) distances.
Note that, here, we’re paying attention to branch lengths.

The optimality score is either the sum of the absolute value of the branch lengths (in the original formulation (Kidd & Sgaramella-Zonta, 1971. Amer. J. Hum. Gen. 23:235).

\[
ME(\tau) = \sum_{k=1}^{2n-3} |v_k|
\]

Or, in more recent formulations (Rhezetsky and Nei, 1992. Mol. Biol. Evol., 9:945), the sum of branch lengths.

\[
ME(\tau) = \sum_{k=1}^{2n-3} v_k
\]

We won’t worry here about how we estimate the \(v_k\)’s. It’s actually very difficult computationally and I teach that in my phylogenetics class.

**D. Maximum Likelihood**

ML estimation, in general, is a standard and incredibly useful statistical procedure.

1) We’re interested in calculating the Likelihood of the observed data given a particular hypothesis.

\[L = \Pr (\text{data} \mid \text{hypothesis})\]

We evaluate several hypotheses, and prefer the one that maximizes the probability of having generated the observed data.
2) Calculation of these probabilities requires an **explicit model**. This has limited its application to molecular data, although there have been recent advances in applying likelihood to morphology.

3) So the application of ML as an optimality criterion in phylogeny estimation is as such:

\[ L(\tau) = \Pr(D \mid \tau) \]

This is simply the probability of the data (the set of aligned sequences), given the tree (and some assumed model of character evolution).

Just as in parsimony, we make the assumption that characters are independent so that we may treat likelihoods for each site in our alignment separately:

\[ L(\tau) = \prod_{i=1}^{s} \Pr(D^i \mid \tau) = \prod_{i=1}^{s} L^i(\tau) \]

where \( s \) is the number of sites (characters) and \( \Pr(D^i \mid \tau) \) is the probability of site \( i \) (character \( i \)) given tree \( \tau \). This is the **single-site likelihood**.

Calculation of the single-site likelihoods is computationally intensive, and again, I go into how it’s done in my phylogenetics class.

**III. Relationships among optimality criteria.**

Lots of the controversy in phylogenetics revolves around choice of optimality criteria, and lots of work (including some of my own) has been focused on performance of the various criteria.

Certainly these differences are important, but often lost in the debate are the various fundamental conceptual similarities that are common to certain sets of them.

Both MP and ML are character based (whereas ME is not).

Both ME and MP minimize the amount of evolution (i.e., sum of branch lengths).

Both ME and ML rely on an explicit model of sequence evolution.
So certainly we should expect there to be a broad range of conditions across which the methods perform similarly, and that certainly has been shown to be the case.

Nevertheless, the differences among them are real, and it’s incredibly informative to study the conditions under which the methods perform differently (i.e., favor different trees).