**Lecture I - Tree-thinking, Parsimony and Cladistics**

**Note: I am updating this course in 2024 to cover less phylogenetics, under the assumption that folks especially interested in phylogenetics will take Anton Suvorov’s graduate course in the spring. However, phylogenetics models are at the core of macroevolutionary trait models, so there is significant overlap. Below, the notes in highlighted in yellow are not required for you to know, but I am keeping them here if you are interested in phylogenetics.**

1. What does a phylogeny represent?
   1. A phylogeny is a map of evolutionary history (onezoom.org). Just as a straight line under relativity may appear bent to us by the effects of gravity, a straight line in evolution appears bent by phylogeny (Metaphor owed to Wayne Maddison & Rich FitzJohn).
   2. **My preferred definition of a phylogeny: A simplified graphical representation of the history of inbreeding populations.**
   3. You must be comfortable with reading and interpreting phylogenies, I have assumed this knowledge, but additional guidance is available in Baum and Smith Chapters 1-4. Be comfortable with:
      1. Terminology: **Monophyletic, paraphyletic, polyphyletic**
      2. **Node, branch/edge, tip/terminal/leaf, root, stem, outgroup**
      3. **Be able to interpret and manipulate a phylogenetic tree.** Understand how it can be displayed graphically, and how relatedness should be interpreted from that the phylogenetic graph.
2. History of Phylogenetic Inference - Introduction
   1. The history of phylogenetic inference and taxonomy/classification (**Systematics**) is simultaneously one of the great scientific success stories and a story of venomous, toxic divisions and tribalism. Remnants of these fights remain, though the younger generation has largely side-stepped these debates, favoring a more pragmatic view. However, understanding the history of progress in phylogenetic inference is quite useful for navigating approaches and the field as a whole.
   2. *Darwin - 1960*’s; ***Evolutionary Taxonomy***. For much of the history of evolutionary biology, systematics was viewed with some skepticism by other fields of evolutionary biology, as it was based largely upon non-reproducible opinions of experts in the field. This is the sort of systematics research that would have been familiar to Darwin: Spend a lifetime working on a group of organisms learning everything you can about their biology, and as an expert, draw informed conjectures regarding their evolutionary history. This approach is largely “model-free” and has been criticized for its lack of objective and reproducible components. On the other hand, it benefits from producing scientists with exceptional knowledge and intuition regarding specific groups of organisms (you will see many of the “old guard” lamenting the loss of such experts).
   3. *Late 1950s - 1960s*; ***Numerical Taxonomy* *& Phenetics***. The 1960s saw the introduction of numerical approaches to systematics. Lead by **Sokal, Sneath and Rohlf**, these researchers believed that inferring the true pattern of common ancestry had no objective methodology, and that we should instead focus on grouping organisms by their overall similarity. Take as many measurements as possible of as many traits as possible, and measure the distance between species. More similar organisms are then clustered together. This group vigorously criticized evolutionary taxonomists.
      1. Methods are often called “**Distance-Matrix**” methods and really marked the beginning of the science of clustering: UPGMA, Neighbor-Joining, Principal Components Analysis, Canonical Variate Analysis.
      2. Decomposes all evolutionary similarity into a single number. One major consequence of this is a loss of information.
      3. Phenetics refers to the *philosophy of classification* applied as much as the actual methods. Thus, many of the arguments are really about interpretations and philosophy more than actual approaches.
   4. *1966 - 1980s*; ***Cladistics and Phylogenetic Systematics****.* In 1966, a German Biologist named **Willi Hennig** introduced a logical framework for inferring phylogenetic relationships that prioritized **derived traits**, which he called “**apomorphies**”, over **ancestral traits** or “**plesiomorphies**”. His followers vigorously attacked the pheneticists as being non-phylogenetic.
      1. The goal of phylogenetic inference and systematics was to identify “monophyletic” (one stem, all descendents) rather paraphyletic (one stem, not all descendents) and polyphyletic (two or more stems) groups.
      2. Estimating groupings based on cladistic principles relies on the ***Principle of Parsimony* or *Occam’s Razor***. Followers of Hennig came to view this as *a priori,* the only philosophically justifiable way to estimate phylogeny.
      3. Difficulty arises from conflicting signal among traits, called ***Homoplasy***. The Maximum Parsimony solution(s) are the phylogenetic trees that minimize the number of changes on the phylogeny.
   5. *1960s - 2000s;* ***Statistical Phylogenetics***. Some systematists began to question the statistical properties and basis of parsimony for evaluating trees. Pioneering work by **Luca Cavalli-Sforza and Anthony Edwards**, and later by **Joe Felsenstein**, introduced the idea that trait evolution on trees could be modeled probabilistically. They showed that parsimony methods can lead to inconsistent estimates of phylogeny under realistic conditions.
      1. These methods are the dominant paradigm today, and encompass the concepts of **Maximum Likelihood** and **Bayesian Inference**.
   6. For much of the 1980s-1990s, systematics conferences were venues for at times tense and at times open warfare between cladists and statistical phylogeneticists. Proponents of the Hennig society argued that Parsimony is the only philosophically valid way to estimate phylogeny, while statistical phylogeneticists argued that theory and simulations showed it inconsistent in many cases. I recommend reading accounts of these meetings, and talking to those who were there. It is an enlightening exercise in understanding how science has progressed (in the past) that I hope will inform how we can better avoid such toxic fights in the future. Eventually, these groups splintered into two distinct societies that now mostly avoid each other, with the Hennig Society running the journal Cladistics and what eventually became the Society for Systematic Biology joining the Society for the Study of Evolution and the American Society of Naturalists in joint “evolution meetings”.
3. Parsimony & Cladistics
   1. **Tree space**. How does one find the most parsimonious tree for a set of traits (a **character matrix**). This is a non-trivial problem for even a small number of taxa. Chapter 3 of Felsenstein covers this information.
      1. The number of rooted, bifurcating, labeled trees is (2n - 3)!/(2n-2(n-2)!)

A tree with 3 taxa has 3 possible topologies

A tree with 5 taxa has 105

A tree with 10 taxa has 34,459,425

A tree with 50 taxa has 2.75292 x 10^76

* 1. Finding the **shortest tree** in this enormous tree space is difficult, and requires computational tools, as there is no analytical solution.
     1. A ***heuristic search*** is a “hill-climbing algorithm”. It is not guaranteed to find the shortest tree, as it may get caught on a ***local peak*** in treespace. It turns out that most phylogenetic algorithms are **reversible**, which means it does not matter where the root of the tree is, and we can search among **unrooted trees,** which reduces the number of trees we must search.
        1. One way to avoid local optima is to have many starting points. The other is to use tree rearrangement algorithms that can traverse valleys in treespace and find **global peaks** more efficiently.
        2. Rearrangement algorithms:
           1. **Nearest Neighbor Interchange (NNI)**
           2. **Subtree pruning and regrafting (SPR)**
           3. **Tree-bisection and reconnection (TBR)**
  2. It turns out there is an algorithm that does guarantee finding the shortest tree, and it is called the **branch and bound algorithm.** 
     1. Essentially, this method discards regions of tree space that it knows cannot have the shortest tree without having to examine them individually. (Felsenstein Chapter 5). Still, even branch and bound method can be computationally intensive if there is substantial homoplasy in the dataset.

1. Parsimony vs. Statistical Phylogenetics
   1. Felsenstein (1979) showed that Parsimony can be an “inconsistent estimator”, meaning that it gets more and more confident in the wrong answer as data are added. The so-called “Felsenstein zone” occurs when long branches on the phylogeny are grouped together due to convergent homoplasy. This “long-branch attraction” also occurs in statistical phylogenetics if the wrong model is specified.
   2. Parsimony has the advantage in that it is more computationally efficient than likelihood or Bayesian methods, and has adherents based on philosophical grounds.
   3. IMO, the major reason statistical phylogenetics is more popular today because it enables more realistic modeling of evolutionary processes and more flexibility. We will cover this more extensively later.
2. Further reading

How to count possible trees: Felsenstein Chapter 3

Heuristic searches: Felsenstein Chapter 4

Branch and bound algorithm: Felsenstein Chapter 5

History of phylogenetic inference: Felsenstein Chapter 10

Parsimonygate: https://www.wired.com/2016/02/twitter-nerd-fight-reveals-a-long-bizarre-scientific-feud/

Joe Felsenstein’s response: [https://www.facebook.com/permalink.php?story\_fbid=937224149693516&id=115402811804635](https://www.facebook.com/permalink.php?story_fbid=937224149693516&id=115402811804635&__xts__%5B0%5D=68.ARCOjtCs9ddjRmn4zqyO2bhFgO0m4id1qDSInpQtKYyvqMQdMiAzRbmCHXxz5hpSr2G841-4S1ZOGTSTggrKdcn4FX5dDl5SZP5zO4lQAUfkU5a3SqRfCXQKhsMSLHz06xzEgzs&__tn__=K-R)