Lecture Outline: Parsimony

1. **Concepts**

   (a) The method of *maximum parsimony* prefers the tree that is consistent with the *most parsimonious* or simplest explanation. For aligned DNA sequences, this is interpreted as the tree that requires the fewest *nucleotide substitutions* to explain the data.

   (b) Most sites of data are not *parsimony informative*. This means that most sites of data can be explained with exactly the same number of substitutions for any tree. For example, an unvaried site with all As can be explained with no changes for any tree. Similarly, a site with an A for one taxon and Gs for all other taxa can also be explained with exactly one substitution for any tree by putting the substitution on the external edge leading to the taxon with an A.

   A site that is *parsimony informative* will be explainable by a minimal number of substitutions for some trees while requiring more substitutions for other trees.

   Example on four-taxon trees with sites AAAA, AGGG, and AAGG.

   (c) The *parsimony score* for each tree is the sum of the smallest number of substitutions needed for each site. The tree with the *lowest parsimony score* is the most parsimonious tree. There are often ties.

   (d) Parsimony does not distinguish between alternative rootings of the same unrooted tree.

2. **Finding a Parsimony Score**

```
           1   4   7  10  13  16  19  22
alligator  GTG AAC TTC CAC --- CGT TGA CTC...
emu        GTG ACA TTC ATT ACT CGA TGA TTT...
kiwi       GTG ACC TTT ACT ACT CGA TGA CTC...
vulture    ATG ACA TTC ATC AAT CGA TGA CTA...
penguin    GTG ACC TTC ATT AAC CGA TGA CTA...
```

   Consider two possible trees: (((E,K),(V,P)),A) and ((((E,V),P),K),A).

   Report score for entire data set.

3. **Fitch Algorithm**

   The algorithm begins with the leaves and works toward the root of the tree. Each node is identified with a subset of the bases. The evaluation is done one site at a time.

   (a) Set the score to be 0.

   (b) Select the next site.

      i. Each leaf is given the set corresponding to its base (or possible bases for gaps and ambiguous characters).

      ii. For each ancestral node whose children have both been processed, let X represent the intersection of the children sets.

      iii. If X is non-empty, the set for the node is X. Otherwise, the set for the node is the union of the children sets and one is added to the score.

      iv. Continue to the root.

   (c) Continue until all sites are processed.

4. **Parsimony Informative Sites**

   (a) A site is parsimony informative if and only if there are at least two pairs of taxa where partners agree with each other but the pairs differ from each other.

   (b) Explain!

5. **Simulation**

Activity 1:

(a) Use dice to simulate DNA on two different four-taxon trees.

(b) Use computer to speed up the process!

(c) Compare the results on the two trees.

6. **Long Branch Attraction**

(a) Long branch attraction is the phenomenon where taxa with long edges tend to be placed together by a method to infer phylogenies.

(b) Parsimony is susceptible to long branch attraction.

(c) Explain!