Maximum Parsimony

Bret Larget

Departments of Botany and of Statistics
University of Wisconsin—Madison

September 13, 2011

Maximum Parsimony

Definition

The *maximum parsimony tree* is the tree that best fits the data in the sense that it can explain the data by requiring the fewest number of changes among all trees. For a given data set, there may be multiple maximum parsimony trees.

Example Data

```
                10              20
                 +               +
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...
```

Example
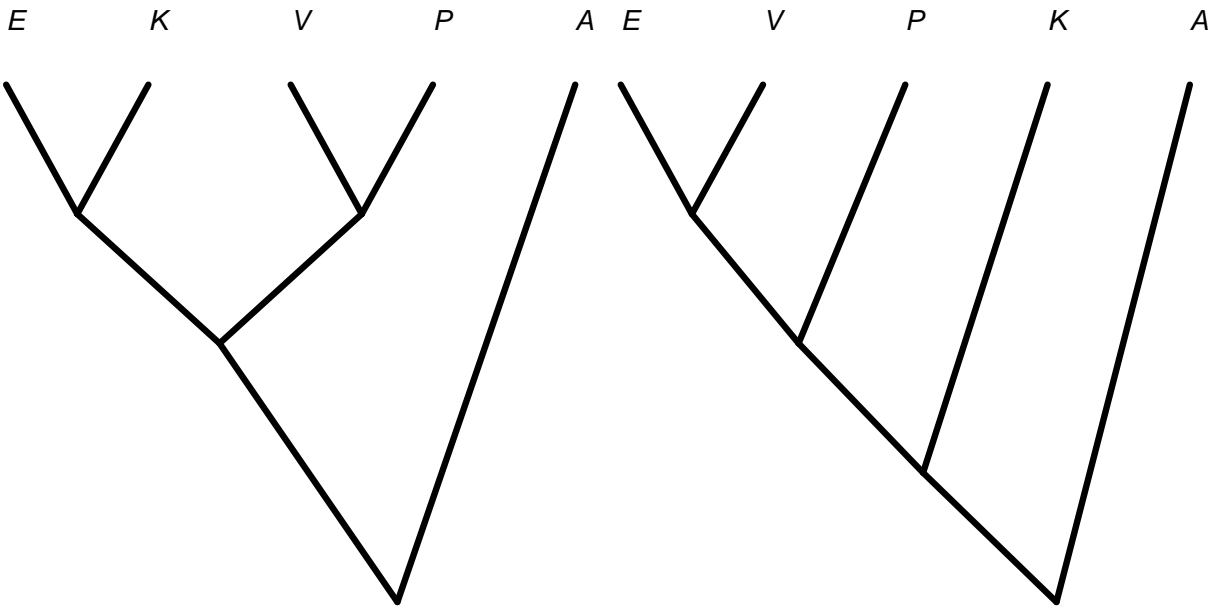
Fitch Algorithm

1. Set the score to be 0.
2. Select the next site.
   1. Each leaf is given the set corresponding to its base (or possible bases for gaps and ambiguous characters).
   2. For each ancestral node whose children have both been processed, let X represent the intersection of the children sets.
   3. 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.
   4. Continue to the root.
3. Continue until all sites are processed.

Simulation Description

- Simulate DNA evolution on two different trees.
- One tree has all short edges.
- One tree has some short edges and two long edges at opposite ends of the tree.
- Pick a base for data at some node (1=A, 2=C, 3=G, 4=T, 5 or 6 = try again).
- For every edge, roll a die to see if there is a change:
  - **short edges:** change when a 1 is rolled
  - **long edges:** change when a 1, 2, or 3 is rolled
- When there is a change, pick a different base at random (1 or 2 = first, 3 or 4 = second, 5 or 6 = third).