

Maximum Parsimony for phylogenetics

(“The shortest possible lecture” - Pete Sheridan)
(“Probably not the true lecture” -Mike Thomas)

MEP: 29 Sept 2005

foundation

- ◆ parsimony methods have been around for about 30 yrs
- ◆ originally developed for morphological data (for which it is very robust)
- ◆ guiding principals:
 - ◆ the tree explained by the smallest number of nucleotide changes is best
 - ◆ the less we need to know about the evolutionary process to the more confidence we can have in our conclusions

assumptions, difficulties

- ◆ no backward or parallel sub's and large n - otherwise, incorrect topology chosen
- ◆ assumes no rate variation among lineages - otherwise, incorrect topology even when n is very large
- ◆ when 2 branches have more subs than other branches in an alignment, they tend to be joined - long-branch attraction (“Felsenstein zone”)
- ◆ very difficult to accurately estimate branch lengths - a fundamental aim in phylogenetics
- ◆ no clear statistical framework to compare trees and test hypotheses - no way to compute means and variances of substitution estimates

strengths

- ◆ clear philosophical framework
- ◆ when divergence is low ($d < 0.1$), rate of substitution is constant, and n is large, parsimony is very good in obtaining the true topology
- ◆ very useful for other types of molecular data (indel events, gene order, etc)

method

- ◆ a nucleotide site is informative only when there are at least 2 different kinds of nucleotides at the site, each of which is represented in at least 2 of the sequences under study
- ◆ for each tree, we calculate (and sum) the minimum number of sub's at each informative site
- ◆ the tree(s) with the fewest sub's is the maximum parsimony tree
- ◆ in the MP tree, informative sites that support the internal branches are synapomorphies, all others are homoplasies
- ◆ the total number of substitutions (informative and uninformative) in a tree is the tree length

Seq \ Site	1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	T	C	A
2	A	G	C	C	G	T	T	C	T
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	T

There are 3 possible trees (unrooted), only sites 5, 7, 9 are informative

some terms

- ◆ invariable site: all seq's have same character
- ◆ variable site: at least one seq has a different character
- ◆ singleton site: variable, but only in one seq
- ◆ parsimony-informative site: at least 2 different nucleotides, each represented in at least 2 sequences
- ◆ homoplasy: informative sites that are inconsistent with the best tree
- ◆ tree length: sum of the minimum number of substitutions for all sites for each topology (may include singleton sites)
- ◆ MP tree: the tree(s) with the smallest tree score
- ◆ weighted MP: some substitutions (e.g., transversions) are assumed to occur at a different rate