Phylogenetic Tree Construction- Character Based Methods- Maximum Likelihood and summary

Manu Madhavan

Lecture 11

Recap

- Phylogenetic Trees - Character based methods
- Maximum Parsimony

Outline

- MP method: Weighted Parsimony, Branch and bound algorithm
- Maximum Likelihood based method

Weighted Parsimony Method

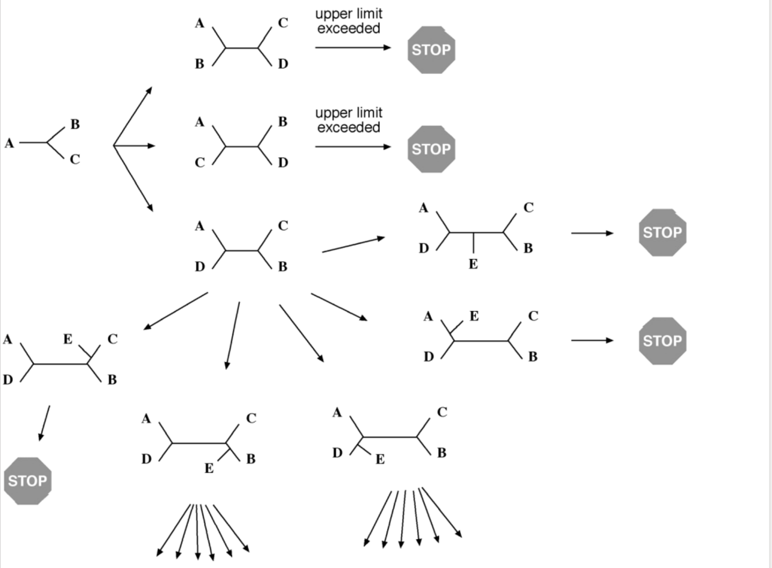
- The parsimony method discussed is **unweighted because it treats all mutations as equivalent**
- Mutations of some sites are known to occur less frequently than others, for example, transversions versus transitions
- A weighting scheme that takes into account the different kinds of mutations helps to select tree topologies more accurately
- Weighting schemes usually come from analyses of empirical data sets

Branch and Bound for Faster Searches

- **Exhaustive search**: the parsimony method examines all possible tree topologies to find the maximally parsimonious tree
- This brute-force approach only works if there are relatively few sequences
- Brute force approach is exponential
- A **branch and bound** method can be used to reduce number of searches

Branch and Bound for Faster Searches

- Establishes an upper limit (or upper bound) for the number of allowed sequence variations.
- Starts by building a distance tree for all taxa involved using either NJ or UPGMA and then computing the minimum number of substitutions for this tree.
- The resulting number defines the upper bound to which any other trees are compared
- Whenever the overall tree length at every single stage exceeds the upper bound, the topology search toward a particular direction aborts
- The rationale is that a maximally parsimonious tree must be equal to or shorter than the distance-based tree


A
B
C
A
C
B
D
upper limit exceeded
STOP
A
B
C
D
upper limit exceeded
STOP
A
C
D
B
A
C
D
B
E
STOP
A
E
C
D
B
STOP
A
E
C
D
B
STOP
A
C
D
E
B
A
C
D
E
B

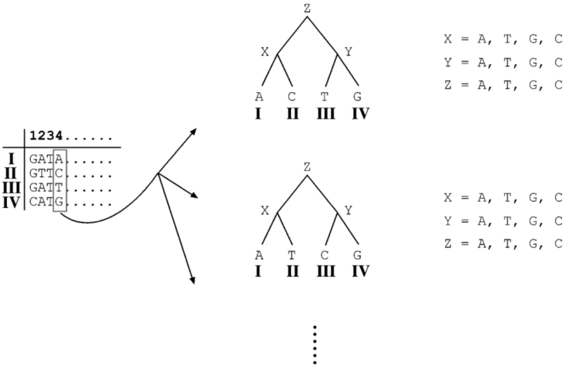
Maximum Likelihood Method

- Probabilistic models to choose a best tree that has the highest probability or likelihood of reproducing the observed
- Exhaustive method that searches every possible tree topology and considers every position in an alignment, not just informative sites
- The likelihood function is calculated from the probability of a residue changing to another residue or remaining the same over the specified period of time
- The likelihood function is calculated for each residue site for each possible tree, and best tree is selected based on the entire sequence

Maximum Likelihood Method

- ML works by calculating the probability of a given evolutionary path for a particular extant sequence
- The probability values are determined by a substitution model (either for nucleotides or amino acids)
- Example: For a DNA sequence, the probability that a nucleotide remains the same after time t is:$P_{ii}(t) = 1/4 + 3/4e^{-\alpha t}$, where where α is the nucleotide substitution rate in the JukesCantor model
- For a nucleotide to change into a different residue after time t, the probability value is determined by $P_{ij}(t) = 1/4 - 1/4e^{-\alpha t}$

Maximum Likelihood Method



```
   |1234......
---+----------
I  |GATA......
II |GTTC......
III|GATT......
IV |CATG......
```

$$L_4 = P(Z \rightarrow X) * P(Z \rightarrow Y) * P(X \rightarrow A) * P(X \rightarrow C) * P(Y \rightarrow T) * P(Y \rightarrow G)$$

$$lnL_4 = ln\ P(Z \rightarrow X) + ln\ P(Z \rightarrow Y) + ln\ P(X \rightarrow A) + ln\ P(X \rightarrow C) + ln\ P(Y \rightarrow T) + ln\ P(Y \rightarrow G)$$

Maximum Likelihood Method

- The overall log likelihood score for a given tree path for the entire sequence is the sum of log likelihood of all individual sites. The same procedure has to be repeated for all other possible tree topologies.
- The tree having the highest likelihood score among all others is chosen as the best tree, which is the ML tree.
- This process is exhaustive in nature and therefore very time consuming
- Quartet Puzzling is an optimization technique

Other approaches

- NJML- hybrid of NJ and ML
- Genetic algorithms

Phylogenetic Trees: Biopython

- `https://biopython.org/wiki/Phylo`