Algorithms as Models of Evolution

Assessing Phylogenetic Signal:

- DNA sequences must be largely free of homoplasy (parallel fixations and reversals).
- Transition/Transversion ratios are greater than for saturated sequences.
- Closely related taxa will have similar Transition/Transversion ratios.
- Saturation of Transitions occurs faster.

Choosing a method requires preliminary assumptions:

- Models of evolutionary process must reflect biological reality.
- NOT all sequences or positions in a sequence will reflect phylogenetic history.
- Ocham's Razor: Simplicity is always preferred (i.e., most parsimonious).
- Errors: Random (or stochastic) and systematic. (More data vs. better assumptions.)

Most methods *share* common assumptions that:

- Nucleotides are evolving **independently** (and are neutrally selected?).
- Comparison involves **orthologous genes** (i.e., divergent evolution).
- **Positional homology** has been inferred correctly.

Estimating Divergence and weighting (before and after the fact):

- Multiple parameter models of nucleotide substitution.
- Must justify weighting rationale: A priori - preconceived notion of evolution?
  A posteriori - estimate observed homoplasy?

The major algorithms to be examined: Remember each is a process!

- Distance Matrix Methods
- Maximum Parsimony
- Maximum Likelihood
- Method of Invariants

Remember: Optimality vs. Single tree issue

Distance Matrix Methods

Distance-based data analysis (as opposed to characters).

UPGMA & WPGMA:

- Gives single tree, not necessarily optimal.
- Groups taxa in order of decreasing similarity.
- Assumes data are Ultrametric (i.e. constant clock). Weakness!
- Both are forms of cluster analysis.

DeSoete Method:

- Yields a single "optimal" tree.
- Does not require ultrametric data.
- Uses a least squares algorithm while satisfying the 4-point condition with a penalty function.

Neighbor-joining:

- Heuristic approach to estimate a single tree with a minimal overall tree length.
- Does not require ultrametric data.
- Does require *additivity assumption*, where the distance between two taxa is equal to the sum of the branches that join them (i.e., 4-point condition).
- Keeps track of nodes not taxa.

Minimum Evolution Method:

- Yields a single "optimal" tree.
- Spirit of parsimony, but uses distance data.
- Searches for the smallest overall length.
- NJ & ME are usually very similar in their resulting tree.

Character-Based Methods

Maximum Parsimony:

- Selects the phylogeny with the minimum number of evolutionary changes (i.e., Ockham's razor).
- Approach relies heavily upon phylogenetically informative characters (i.e., those with two or more states shared by two or more taxa).
- Popularity is due to logical simplicity.
- Permits unequal rates and assumes homoplasy is minimal.
- Minimizes total tree length, or the number of steps required to explain a given data set.

Maximum Likelihood:

- Calculates probability of data set, given a particular model of evolutionary change.
- Independently calculates probabilities at each site, with joint least probability for tree.
- Frequently the method least affected by sampling error.
- Likelihood is proportional to the probability of the data given the tree; it is *NOT* the probability of the tree given the data.
- More robust to systematic errors.

Method of Invariants:

- Counts the number of transversion events supporting phylogeny after adjusting for homoplastic change.
- Designed for 4-taxon problems where homoplasy is expected to be high (e.g., distantly related taxa with unequal rates of evolution).
- Requires a balance between specific classes of transvertions.
- Relatively long sequences are required, therefore inefficient.
- No assumption about rates!

Resampling techniques – Primarily through Bootstrapping:

- Characters randomly drawn with replacement, leading to new data set of original size.
- Express with consensus tree.