

Distance methods for tree construction & evaluation

MEP: 5 Oct 2004

tree types (from last week)

- ◆ phenogram: branching diagram that links entities by estimates of overall similarity (usually based on cluster analysis)
- ◆ cladogram: depicts historical branching relationships among entities; branch lengths are arbitrary - only order is significant
- ◆ phylogram: depicts historical branching relationships among entities; branches are proportional to the amount of inferred character change

problems with distance methods for phylogenetics

- ◆ lost information: when nucleotide sites are converted into a distance matrix, we can't trace the evolution of individual sites of classes of sites on the tree
- ◆ branch lengths are conceptually difficult to interpret: if a tree has a length of 331.5 substitutions, what's "half of a subst?"

UPGMA

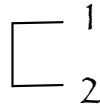
- ◆ PAUP command: "begin paup; upgma; end;"
- ◆ "unweighted pair-group method using arithmetic averages" - Sneath & Sokal (1973)
- ◆ creates a phenogram: represents phenetic similarity for a group of species
- ◆ when rates of gene substitution is constant, it is fairly good at making reconstructing a phylogeny
- ◆ easily, logically rooted, but not always in the correct spot...

given a distance matrix:

Taxon	1	2	3	4
2	d_{12}			
3	d_{13}	d_{23}		
4	d_{14}	d_{24}	d_{34}	
5	d_{15}	d_{25}	d_{35}	d_{45}

where d_{ij} is the distance from taxon i to taxon j ,
 computed using an appropriate distance correction
 and d_{12} is the smallest distance in the matrix

Tree:

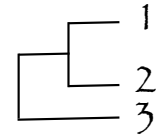


given a distance matrix:

Taxon	$u=(1+2)$	3	4
3	d_{u3}		
4	d_{u4}	d_{34}	
5	d_{u5}	d_{35}	d_{45}

where d_{uj} is the distance from clade $(1+2)$ to taxon j ,
 and d_{u3} is the smallest distance in the matrix

Tree:

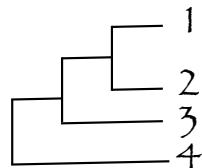


given a distance matrix:

Taxon	$v=(1+2+3)$	4
4	d_{v4}	
5	d_{v5}	d_{45}

where d_{vj} is the distance from clade $(1+2+3)$ to taxon j ,
 and d_{v4} is the smallest distance in the matrix

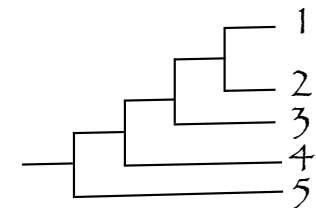
Tree:



given a distance matrix:

- the last taxon is added and the root is placed
- branch length are equal within a clade, such that $d_{uk} = (d_{1k} + d_{2k})/2$, where k is another taxon (3, 4, or 5)
- for any 2 clusters (A & B), $d_{AB} = \sum d_{ij} / (rs)$, for all values of ij where r & s are the numbers of taxa in clusters A & B, respectively, and d_{ij} is the distance between taxon i in cluster A and taxon j in cluster B

Tree:



neighbor-joining (NJ)

- ◆ PAUP: "begin paup; nj; end;"
- ◆ designed to be an efficient approximation of the minimum evolution and maximum likelihood methods of finding the tree with the least amount of evolution change
- ◆ neighbors: taxa connected by a single node in an unrooted tree
- ◆ begins with a star tree (no neighbors) - notice similarity to exhaustive search, which starts with a core tree and successively adds taxa:

NJ:

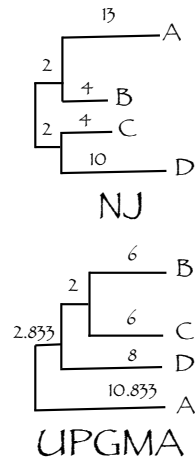
- ◆ basically, we take shortcuts to find a tree that minimizes the sum of all branch lengths, $S = \sum b_i$, where b_i is the length of branch i
- ◆ we begin with a star tree and at each step find the neighbors that, when joined, causes the largest reduction in S
- ◆ we are finished when all taxa are joined in a single unrooted tree

- ◆ for the star tree, $S_0 = \sum L_{iX}$, the sum of all branch lengths to the central node (X)
- ◆ for the tree in which 1 and 2 are joined as neighbors:
 - ◆ $S_{12} = [(2T - R_1 - R_2) / 2(m-2)] + [d_{12} / 2]$, where T is $\sum d_{ij}$, R_1 is $\sum d_{1i}$ (for $i \Rightarrow m$), and R_2 is $\sum d_{2i}$ (for $i \Rightarrow m$)
- ◆ the S_{ij} can be computed for any tree replacing 1,2 with i,j

- ◆ branch length from a new node A to newly joined taxa i,j :
 - ◆ $b_{Ai} = [1/2(m-2)] / [(m-2)d_{ij} + R_i - R_j]$
 - ◆ $b_{Aj} = [1/2(m-2)] / [(m-2)d_{ij} - R_i + R_j]$
- ◆ distance from A to any other taxon k :
 - ◆ $d_{Ak} = (d_{ik} + d_{jk} - d_{ij}) / 2$

NJ v. UPGMA

Taxon	A	B	C	D
A	0			
B	17	0		
C	21	12	0	
D	27	18	14	0



Nei's advice on selecting a distance measure for phylogeny reconstruction (ModelTest approach is generally superior):

- when $JC\ d < 0.05$, use JC or p even if there is Ts/Tv bias or variation in r among sites
- when $0.05 < d < 1.0$ and:
 - n is small, use p
 - n is large, use JC
 - n is large and there is Ts/Tv bias ($R > 5$), use K2P
 - n is v. large and there is much variation among sites, use HKY+ Γ
- when $d > 1.0$, eliminate fast evolving portions of gene or simply avoid...

4. for closely related species, consider using d_S (of protein-coding genes); for distantly related species, consider using d_N or amino acid sequences

5. if two distances give similar values, use the simpler one because it has a smaller variance

Note: these guidelines are for picking distances for tree reconstruction - for estimating branch lengths or evolutionary times, always use unbiased estimators

Minimum evolution:

- when an unbiased distance estimator is used, the value of S is smallest for the true tree irrespective of the number of sequences used (m)
- when m is relatively small, the the ME tree is generally identical or close to the NJ tree - one strategy to find the ME tree is to start with the NJ tree and do some local rearrangements (a heuristic ME search)
- alternatively, we can look at every tree and compare pairs of trees, testing the hypothesis $D = S_B - S_A$; for a statistical test, we need to know the expected error

Least-Squares (LS) fit

- ◆ LS measures the sum of the difference between:
 - ◆ pairwise distances
 - ◆ tree distances for pairs (sum of branch lengths between a pair of sequences)
- ◆ this difference indicates the quality of a given topology
- ◆ the topologies are examined and the one(s) with the best LS scores are chosen