Categories of tree reconstruction methods

	Parsimony	Distance	Maximum likelihood estimation	Bayesian methods
Character data	x		x	x
Pairwise distances		x		

Distance-based methods

- Obtaining a distance matrix from an alignment and correcting for multiple substitutions
- Fitting distances to a tree

Nov 29th

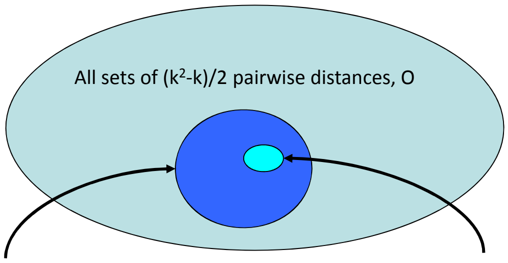
  - – Conditions for obtaining an exact fit
    - Additive distances
    - Ultrametric distances

Dec 1st

  - – Greedy algorithms
    - UPGMA
    - NeighborJoining

Dec 6th

Overview



Additive distances

- The four point condition holds for all quartets in O
- The distances fit a unique unrooted tree
- Branch lengths and topology are uniquely determined.
- ➢ A greedy algorithm finds this tree in polynomial time: **Neighbor Joining**

Ultrametric distances

- The three point condition holds for all triples in O
- The distances fit a unique *rooted* tree; *all leaves are equidistant from the root.*
- Branch lengths and topology are uniquely determined.
- ➢ A greedy algorithm finds this tree in polynomial time: **UPGMA**

Summary

- A matrix is *additive* if it satisfies the four point condition.
- A tree defines a *tree metric, T[i,j]; i.e.,* the pairwise distances between all pairs of leaves.
- All tree metrics are additive.
- If a matrix, *O[i,j],* is additive
  - there exists a unique tree topology with branch lengths such that *T[i,j]* = *O[i,j].*
  - This tree can be obtained in polynomial time.
- In real life, observed distance matrix, *O[i,j]* is never additive.

Summary, cont’d

- A matrix is *ultrametric* if it satisfies the three point condition.
- All ultrametric matrices fit rooted trees.
- Not all rooted tree metrics are ultrametric.
- An ultrametric tree
  - satisfies the molecular clock hypothesis.
  - All distances from the root to a leaf are the same.
  - Its branch lengths are proportional to time.
- For $k > 3$,
  - All ultrametric matrices are additive
  - But, an additive matrix is *not necessarily* ultrametric.

Distance-based methods

- Obtaining a distance matrix from an alignment and correcting for multiple substitutions
- Fitting distances to a tree
  - Conditions for obtaining an exact fit
    - Additive distances
    - Ultrametric distances
  - ➢ Greedy algorithms
    - ➢ UPGMA
    - NeighborJoining

Greedy methods for distance-based phylogeny reconstruction

Taxa are points in a metric space with pairwise distances, *D[i,j]*. Tree building is equivalent to hierachical clustering of these points.

These greedy algorithms maintain a forest of subtrees, beginning with the set of singleton trees (i.e., trees with one leaf and no edges). At each iteration, the algorithm merges two neighboring subtrees in the forest. The length(s) of edge(s) connecting the subtrees are calculated and the distance matrix is updated. This step is repeated until only one tree remains - the final result.

The algorithms differ in

- How neighbors to be merged are identified.
- How the branch lengths are computed.
- How the distance matrix is updated.

Unweighted Paired Group Method with Average Means (UPGMA)

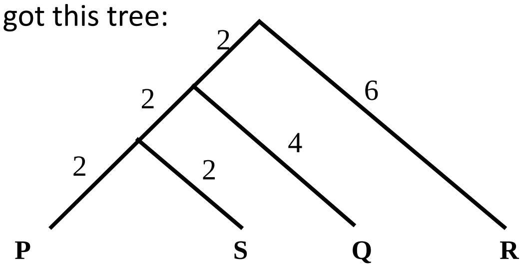
- The UPGMA algorithm is a variant of average linkage. UPGMA is based on the molecular clock assumption. The consequences of this assumption are that
- At each step, the two closest taxa are selected as neighbors.
- The height of the least common ancestor of any pair of leaves is half the distance between the leaves.
- If a distance matrix, *D[]*, is ultrametric, then UPGMA will reconstruct the correct rooted tree in quadratic time.

UPGMA: An example

Observed distances:

	Q	R	S
P	9	9	4
Q	0	16	7
R		0	11

In class, we applied UPGMA to
this distance matrix and got this tree:



UPGMA: An example

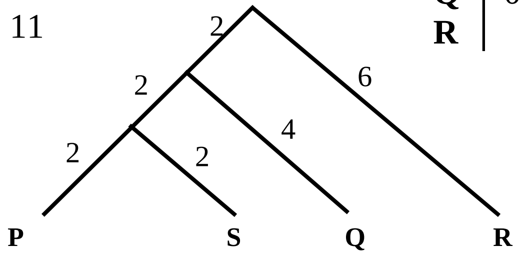
Note, however, that the tree distances are very different from the observed distances.

Observed distances:

O	Q	R	S
P	9	9	4
Q	0	16	7
R		0	11

Tree distances:

T	Q	R	S
P	8	12	4
Q	0	12	8
R		0	12



UPGMA: An example

We obtained the wrong tree because *O* is not ultrametric!

Consider the triple, {P,Q,R}

- *9 ≤ max(9,16)*
- *9 ≤ max(9,16)*
- *16 ≰ max(9.9)*

O	Q	R	S
P	9	9	4
Q	0	16	7
R		0	11

UPGMA: An example

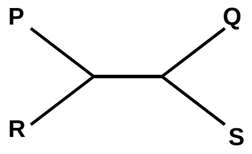
However, the matrix is additive:

PQ +RS = 29+11=20

PR +QS = 9+7=16

PS +QR = 4+16=20

From the pairwise distances, we know that the topology of this tree must be:



which is not the topology we obtained with UPGMA.

O	Q	R	S
P	9	9	4
Q	0	16	7
R		0	11

UPGMA: Another example

As an exercise, verify that

1. this matrix is ultrametric and
2. when UPGMA is applied to this matrix, you obtain the correct tree

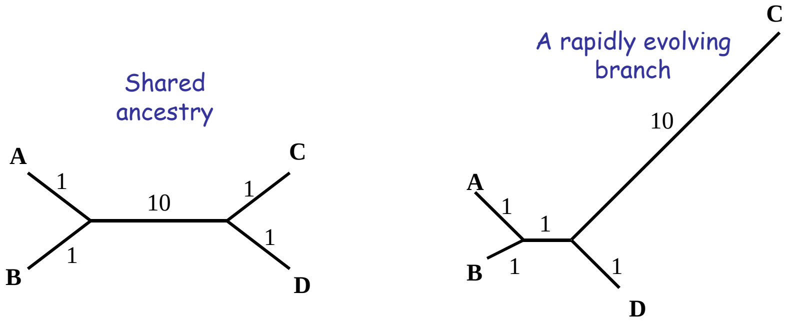
T	**Q**	**R**	**S**
P	8	12	4
Q	0	12	8
R		0	12

Neighbor Joining

The NJ algorithm adjusts the distance matrix for variations in the rate of change. The “adjusted" distance between a pair of nodes is calculated by subtracting the average of the distances to all other leaves.

- **Thm:**
  - If *D* is additive, the pair of taxa that minimimize this "corrected" distance matrix are neighbors in the true tree.
- **Proof:**
  - Durbin *et al.*, 7.8
- If *D* is additive, then NJ will reconstruct the correct *unrooted* tree in quadratic time.

NJ intuition



Does a long branch indicate shared ancestry or a change in the substitution rate?

Neighbor Joining example

In class, we worked through one iteration of the NJ algorithm on this matrix, and verified that:

1. NJ does correctly select *(P, R)* and (*S, Q)* as neighbors and
2. NJ correctly calculated the branch lengths from P and R to their common ancestor.

O	**Q**	**R**	**S**
P	9	9	4
Q	0	16	7
R		0	11

On your own, work through the NJ algorithm to complete reconstruction of this tree to convince yourself that NJ does, indeed, correctly reconstruct trees from additive matrices.