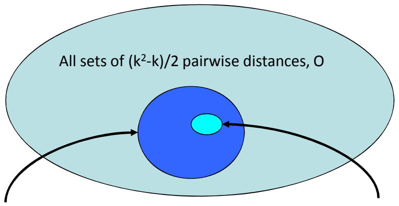
Categories of tree reconstruction methods

	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
    - Additive distances
    - Ultrametric distances
  - ➢ Greedy algorithms
    - UPGMA
    - NeighborJoining

Overview



Additive distances

- 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 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.
- A matrix is *additive* if it is a tree metric.
- 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

- 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.
- A matrix is *ultrametric* if it fits an ultrametric tree.
- 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)

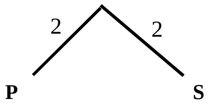
- The UPGMA algorithm is a variant of average linkage. UPGMA is based on the molecular clock assumption. The consequences of this assumption are that
- At each step, the two closest taxa are selected as neighbors.
- The height of the least common ancestor of any pair of leaves is half the distance between the leaves.
- If a distance matrix, *D[]*, is ultrametric, then UPGMA will reconstruct the correct rooted tree in quadratic time.
- Pseudocode for the UPGMA algorithm is linked to the course website: http://www.cs.cmu.edu/~durand/03-711/Notes/upgma04.pdf

UPGMA: An example

Observed distances:

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

- UPGMA selects the closest pair of nodes and makes them neighbors in the tree with height = distance/2.
- This assumption is based on the assumption that the rate of change is the same in all lineages.

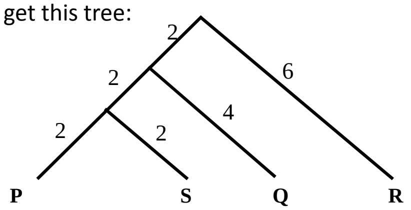


UPGMA: An example

Observed distances:

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

If you apply UPGMA to the above distance matrix, you will get this tree:



UPGMA: An example

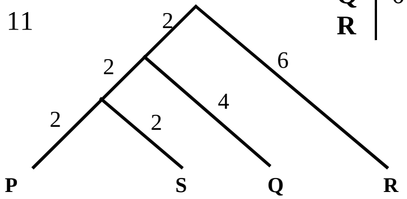
Note, however, that the tree distances are very different from the observed distances.

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}

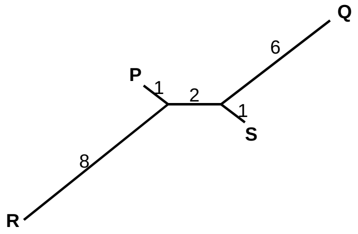
- *9 ≤ max(9,16)*
- *9 ≤ max(9,16)*
- *16 $\cancel{\leq}$ max(9.9)*

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

UPGMA: An example

However, the matrix is additive:

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

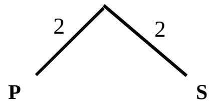


UPGMA: Another example

Observed distances:

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

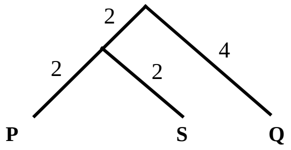
- UPGMA selects the closest pair of nodes and makes them neighbors in the tree with
  height = distance/2.
- This assumption is based on the assumption that the rate of change is the same in all lineages.



UPGMA: Another example

T	Q	R
P,S	8	12
Q	0	12

- Recalculate the distances , treating P and S as a single entity, and select the next pair of neighbors.

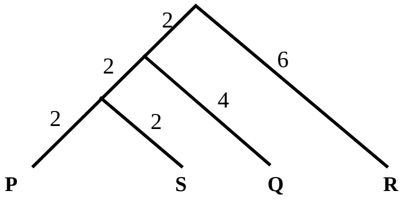


UPGMA: Another example

Observed distances:

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

The final result:

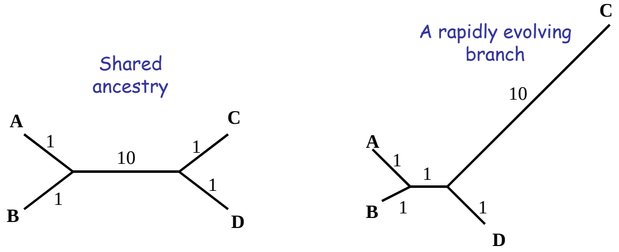


Neighbor Joining

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.
- Pseudocode for the NJ algorithm is linked to the course website: http://www.cs.cmu.edu/~durand/03-711/Notes/nj04.pdf

NJ intuition



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