Unweighted Pair Group Method with Average Means

Algorithm: UPGMA
Input:
A matrix O of observed pairwise distances on k taxa.
Initialization:
Assign each taxon i to its own cluster C_i.
Let $\mathcal{T} = t_1, \ldots, t_k$ be the set of subtrees with one leaf.
For all $1 \leq i, j \leq k$, let $D[i,j] = O[i,j]$
Iteration:
While ($|\mathcal{T}|$ > 1)
{
Find two taxa i and j such that $D[i,j]$ is minimal.
Create a new subtree t_l with root l such that
l is the parent of i and j
height $(l) = D[i,j]/2$
Define a new cluster $C_l = C_i \cup C_j$.
For all $m \neq i, j$,
Compute $D[l,m]$ according to Equation 1.
Remove t_i and t_j from $\mathcal{T}$ and add t_l
}

The distance between node l, with children i and j, and node $m \neq i, j$ is

$$D[l,m] = \frac{|C_i| \cdot D[i,m] + |C_j| \cdot D[j,m]}{|C_i| + |C_j|}. \quad (1)$$

This is a modified version of the algorithm given in Durbin *et al.* Note that D is expanded to include the pairwise distances between all nodes, whether internal nodes or leaves.

The distance between nodes u and v representing clusters C_u snd C_v can be also be calculated using distances between leaf nodes only:

$$D[u,v] = \frac{1}{|C_u| \cdot |C_v|} \sum_{i \in Cu} \sum_{j \in Cv} D[i,j].$$

Dr. Singh's lecture notes use this approach.