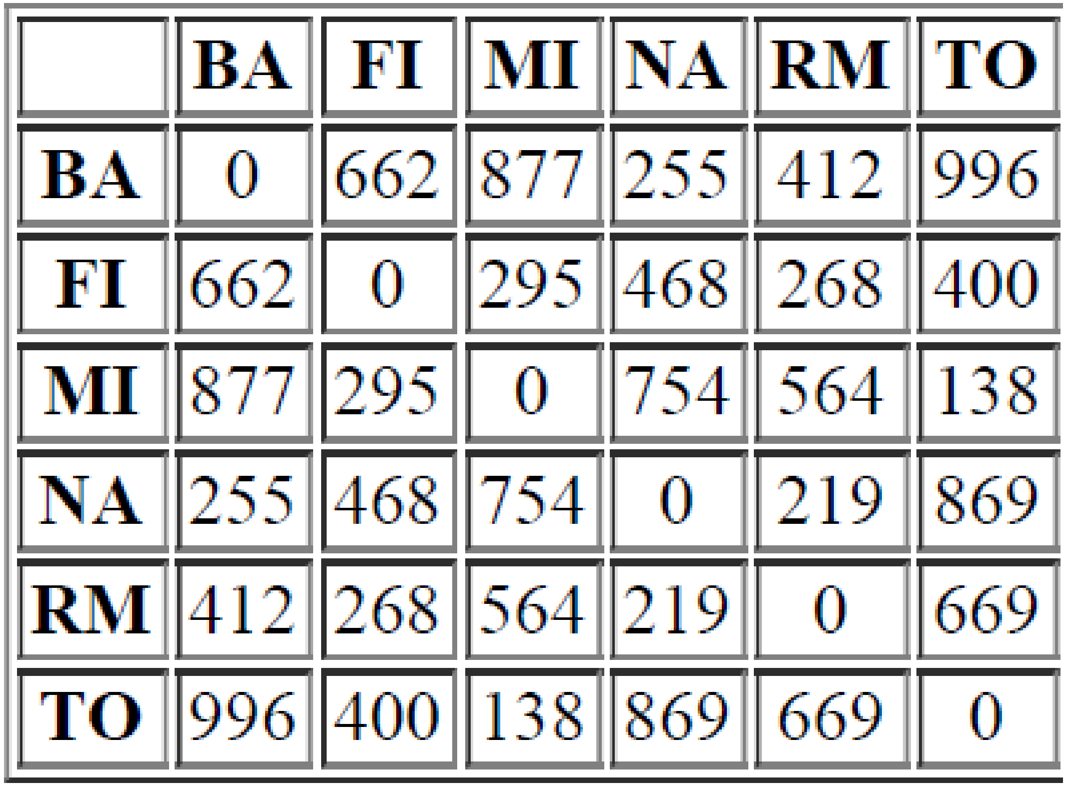
SingleLinkage Clustering: The Algorithm

Lecture Four

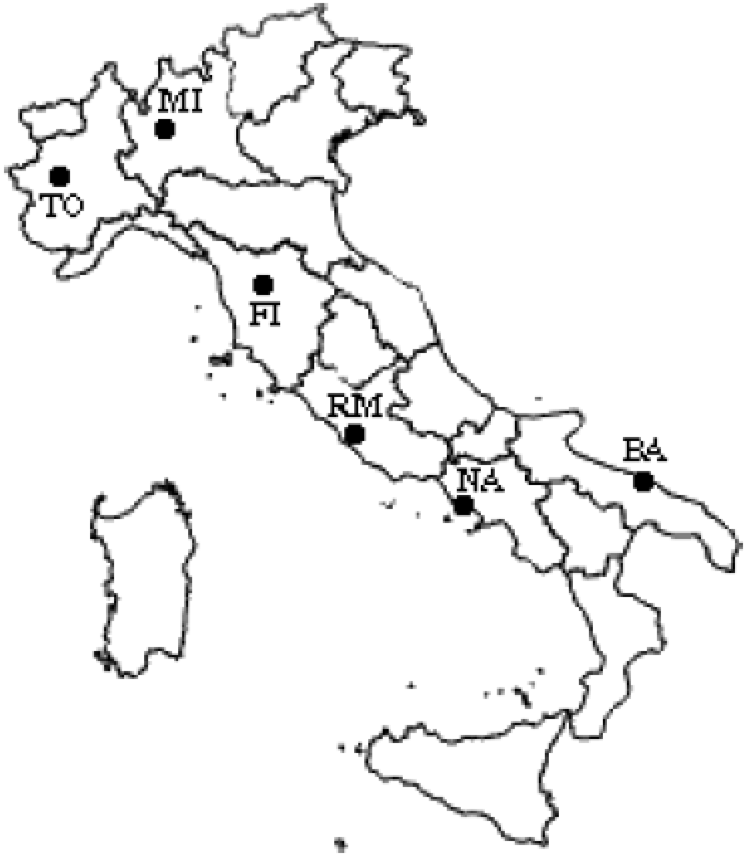
Example

- Let’s now see a simple example: a hierarchical clustering of distances in kilometers between some Italian cities.
- The method used is singlelinkage.
- Input distance matrix (L = 0 for all the clusters):

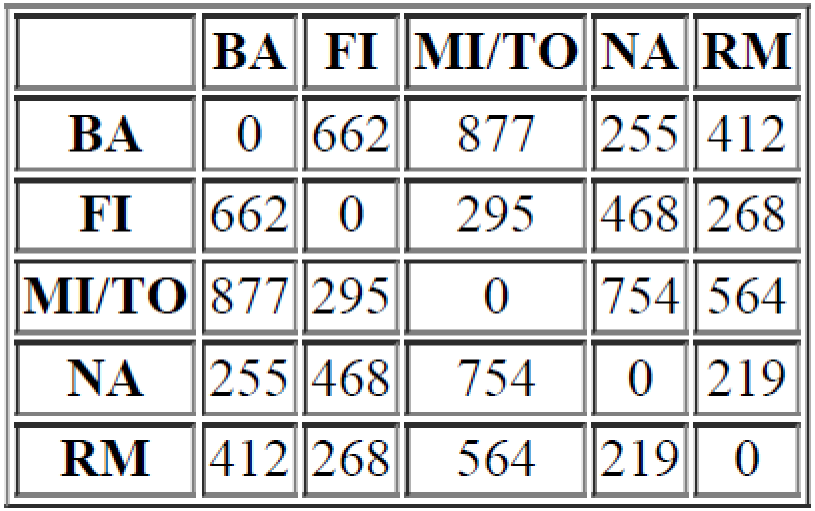
Step 1



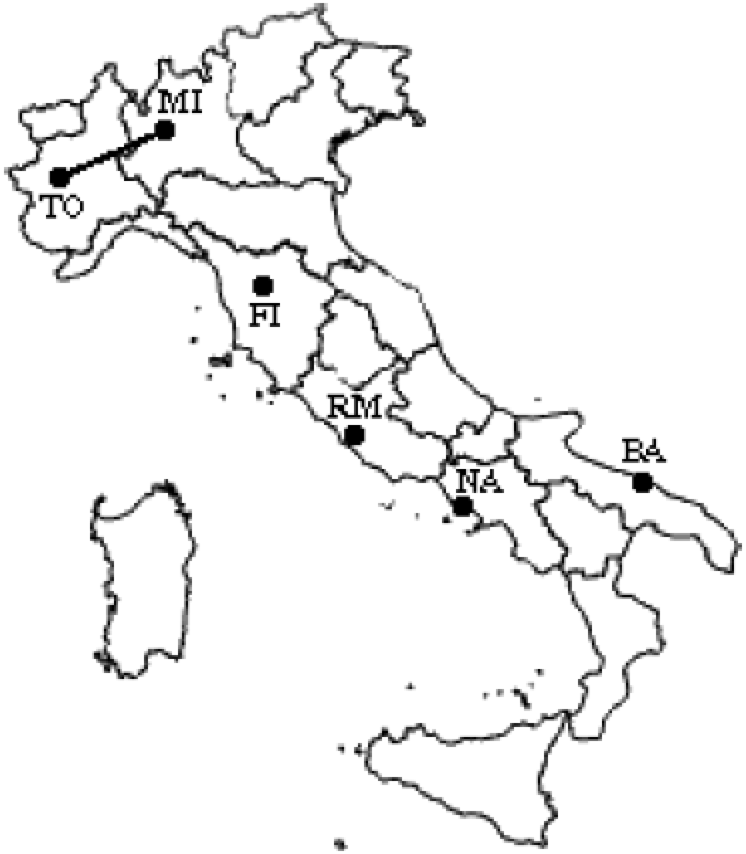
	BA	FI	MI	NA	RM	TO
BA	0	662	877	255	412	996
FI	662	0	295	468	268	400
MI	877	295	0	754	564	138
NA	255	468	754	0	219	869
RM	412	268	564	219	0	669
TO	996	400	138	869	669	0



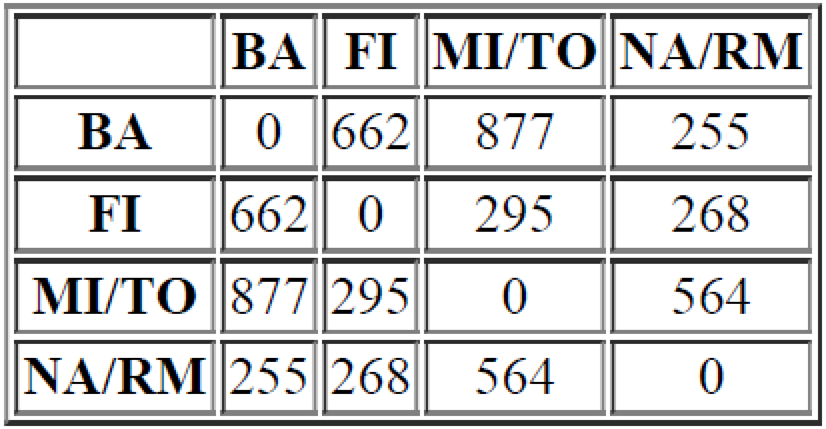
Step 2



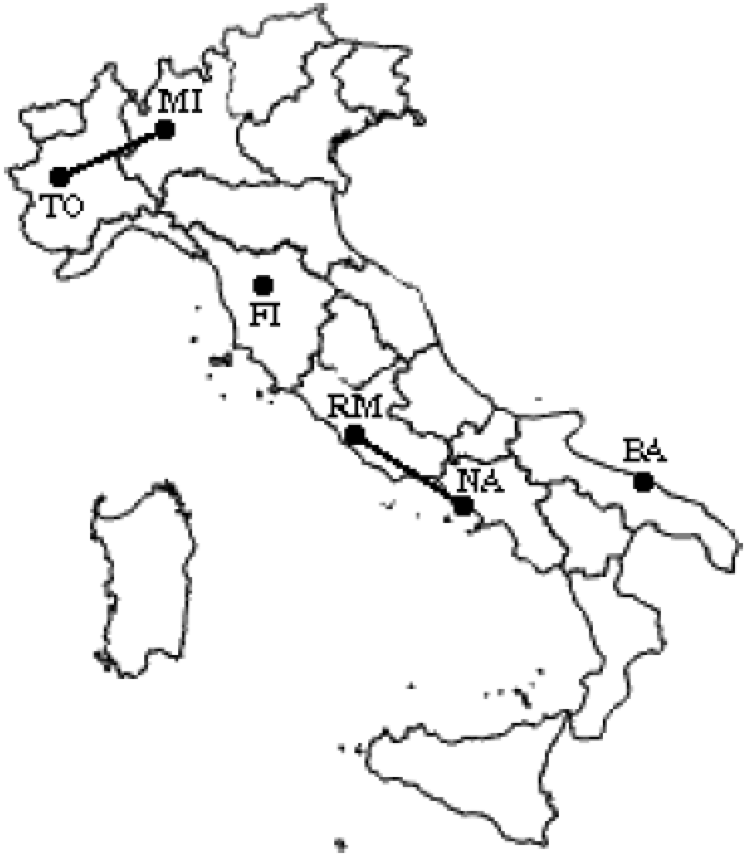
	BA	FI	MI/TO	NA	RM
BA	0	662	877	255	412
FI	662	0	295	468	268
MI/TO	877	295	0	754	564
NA	255	468	754	0	219
RM	412	268	564	219	0



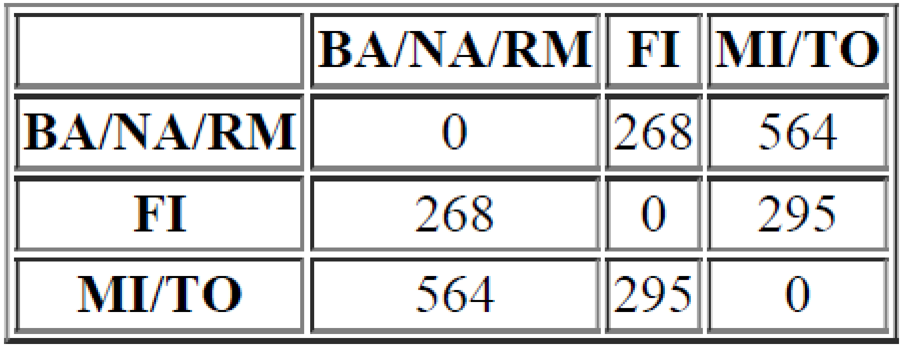
Step 3



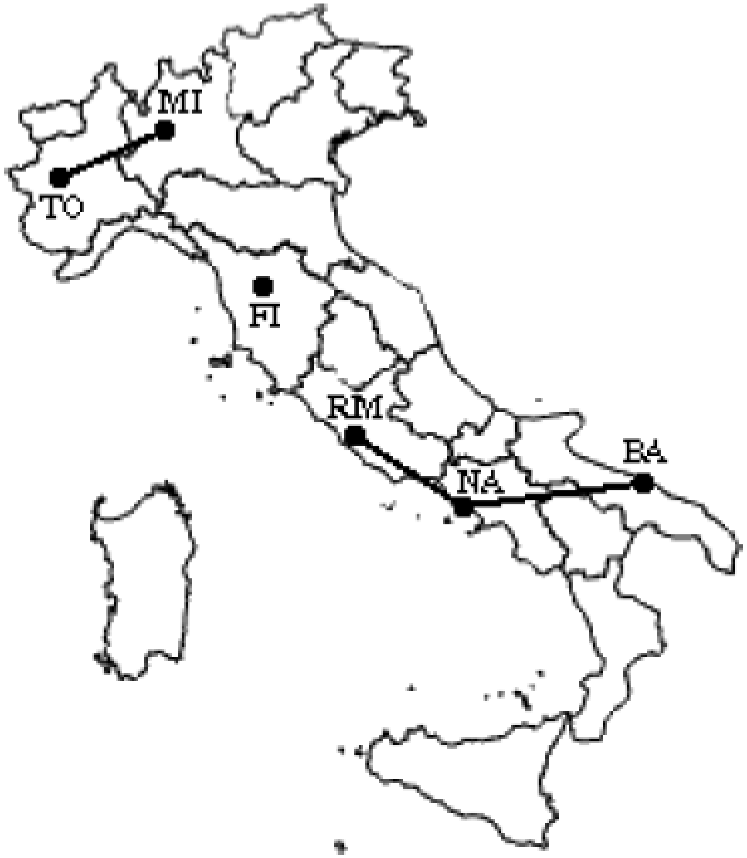
	BA	FI	MI/TO	NA/RM
BA	0	662	877	255
FI	662	0	295	268
MI/TO	877	295	0	564
NA/RM	255	268	564	0



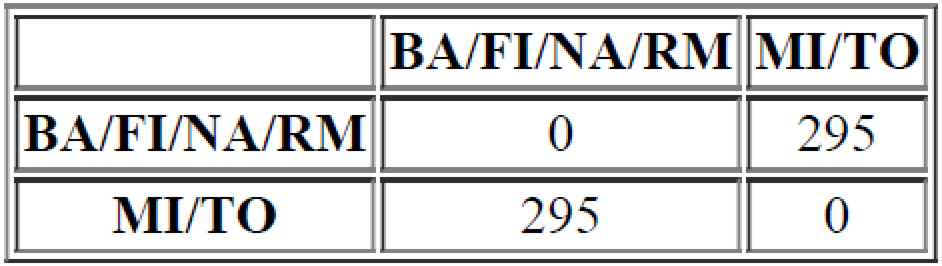
Step 4



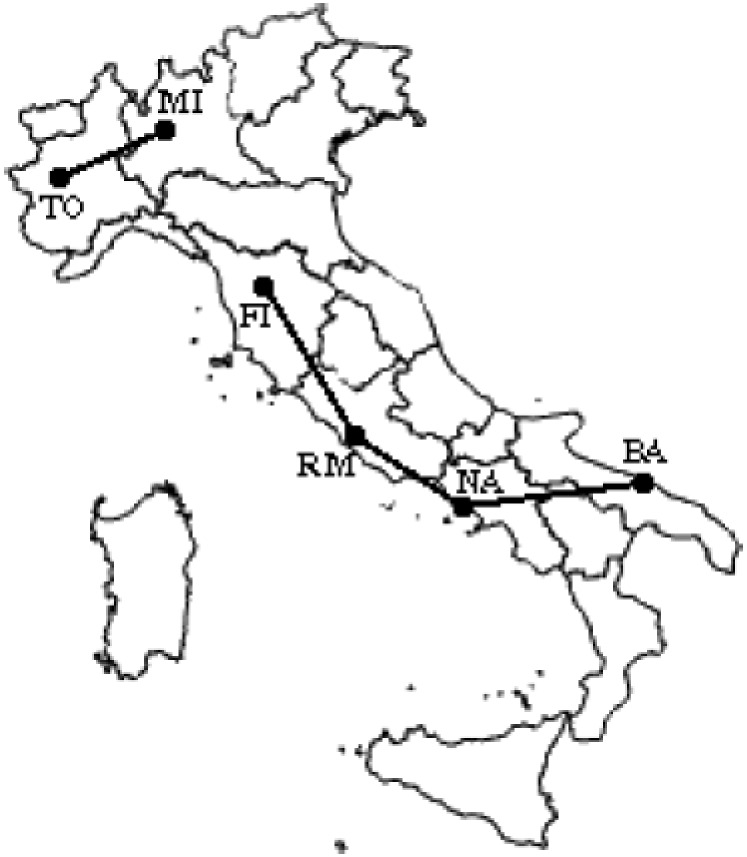
	BA/NA/RM	FI	MI/TO
BA/NA/RM	0	268	564
FI	268	0	295
MI/TO	564	295	0



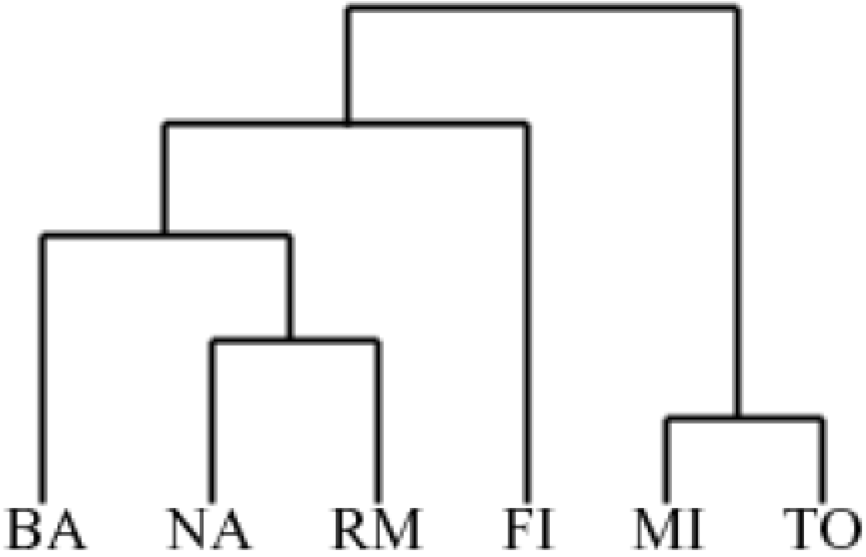
Step 5



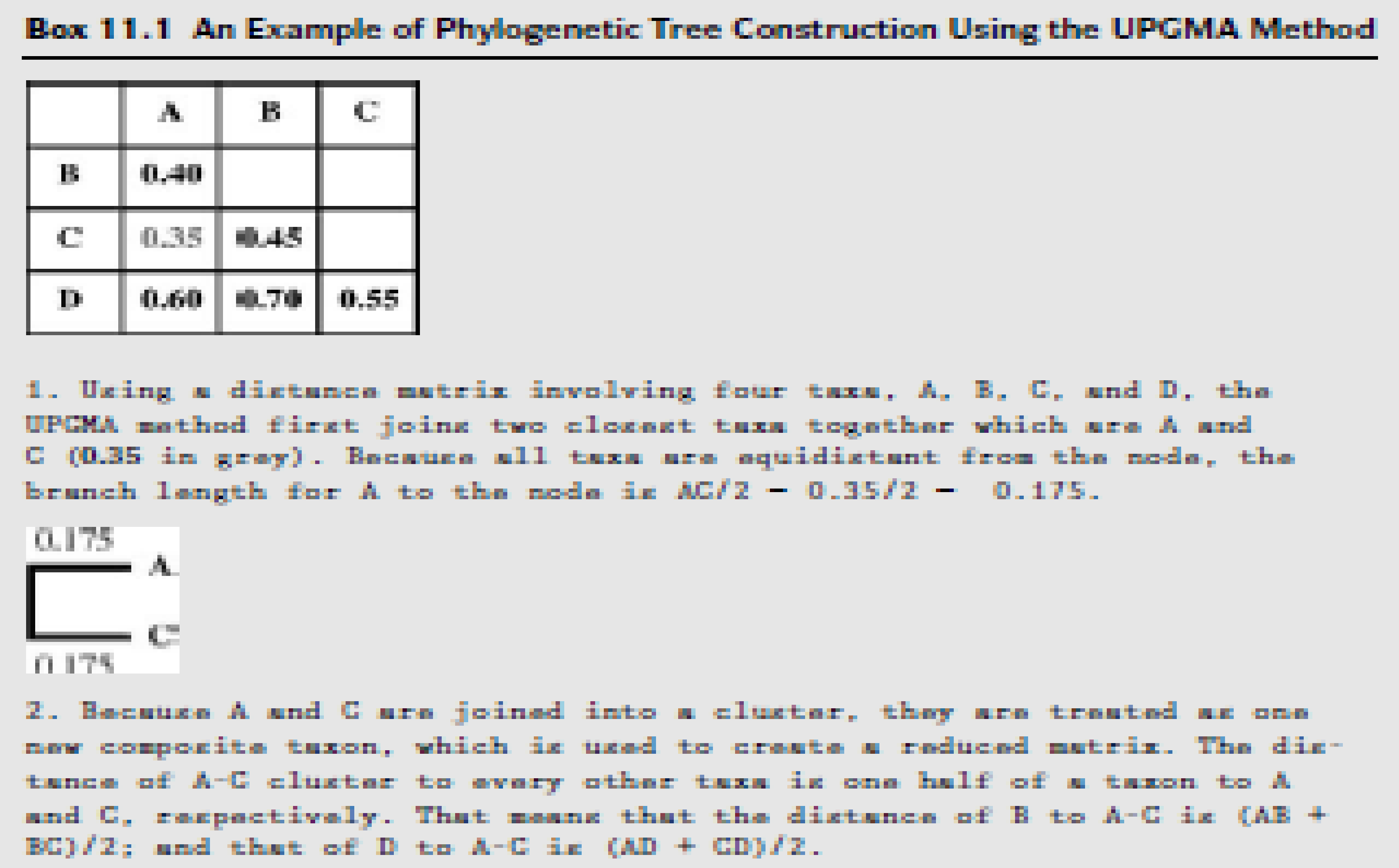
	BA/FI/NA/RM	MI/TO
BA/FI/NA/RM	0	295
MI/TO	295	0



Final



Biological Example



Box 11.1 An Example of Phylogenetic Tree Construction Using the UPGMA Method

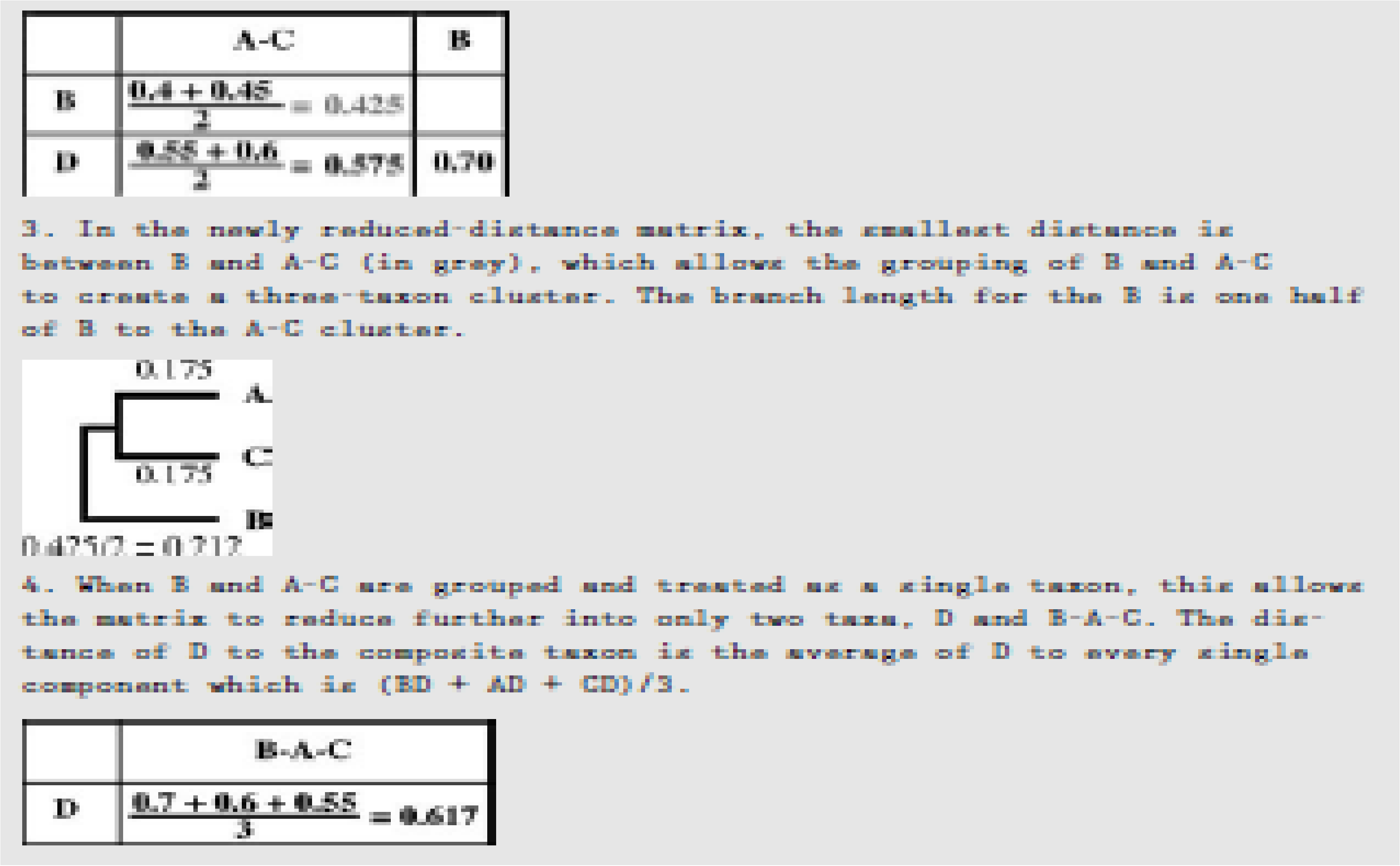
	A	B	C
B	0.40		
C	0.35	0.45	
D	0.60	0.70	0.55

1. Using a distance matrix involving four taxa, A, B, C, and D, the UPGMA method first joins two closest taxa together which are A and C (0.35 in gray). Because all taxa are equidistant from the node, the branch length for A to the node is AC/2 = 0.35/2 = 0.175.

0.175
A
C
0.175

2. Because A and C are joined into a cluster, they are treated as one new composite taxon, which is used to create a reduced matrix. The distance of A–C cluster to every other taxa is one half of a taxon to A and C, respectively. That means that the distance of B to A–C is (AB + BC)/2; and that of D to A–C is (AD + CD)/2.

Biological Example (Cont.)



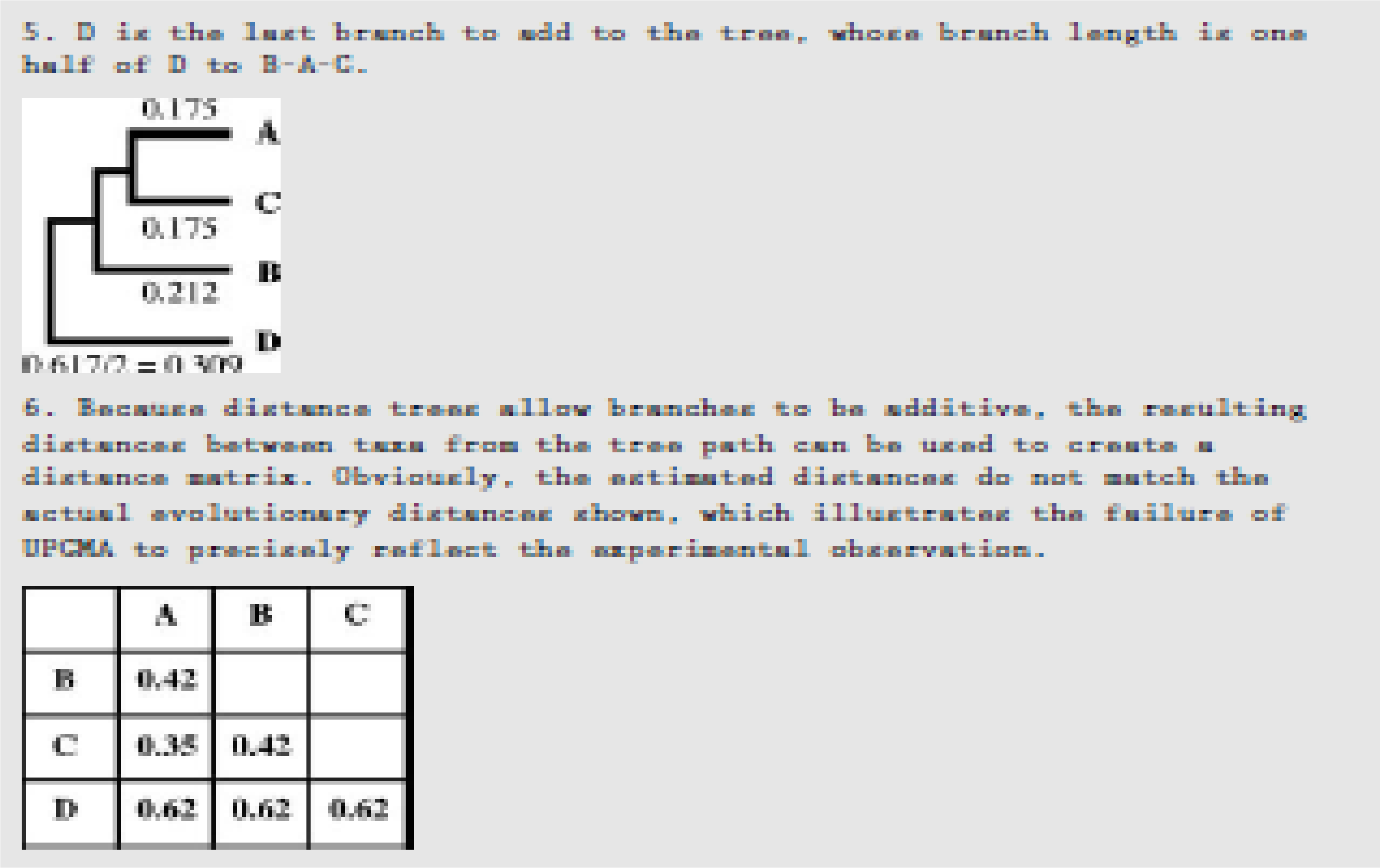
	A-C	B
B	$\frac{0.4 + 0.45}{2} = 0.425$	
D	$\frac{0.55 + 0.6}{2} = 0.575$	0.70

3. In the newly reduced-distance matrix, the smallest distance is between B and A-C (in gray), which allows the grouping of B and A-C to create a three-taxon cluster. The branch length for the B is one half of B to the A-C cluster.

4. When B and A-C are grouped and treated as a single taxon, this allows the matrix to reduce further into only two taxa, D and B-A-C. The distance of D to the composite taxon is the average of D to every single component which is (BD + AD + CD)/3.

	B-A-C
D	$\frac{0.7 + 0.6 + 0.55}{3} = 0.617$

Biological Example (Cont.)



5. D is the last branch to add to the tree, whose branch length is one half of D to B-A-C.

6. Because distance trees allow branches to be additive, the resulting distances between taxa from the tree path can be used to create a distance matrix. Obviously, the estimated distances do not match the actual evolutionary distances shown, which illustrates the failure of UPGMA to precisely reflect the experimental observation.

	A	B	C
B	0.42		
C	0.35	0.42	
D	0.62	0.62	0.62