Problem 5.1

Chen Bo Calvin Zhang

02/11/2020

This problem will explore a number of hierarchical clustering methods on the UCI Zoo data set. This data set contains measurement of 16 features on 101 animals, and we use cluster analysis to find groups of similar animals.

```
library("mlbench")
data(Zoo)

X.zoo = Zoo[,1:16]
L.zoo = factor(Zoo[,17])

print(dim(X.zoo)) # 101 samples
```

```
## [1] 101  16
```

```
# 16 features
print(colnames(X.zoo))
```

```
##  [1] "hair"     "feathers" "eggs"     "milk"     "airborne" "aquatic"
##  [7] "predator" "toothed"  "backbone" "breathes" "venomous" "fins"
## [13] "legs"     "tail"     "domestic" "catsize"
```

```
# seven zoological groups
print(levels(L.zoo))
```

```
## [1] "mammal"        "bird"          "reptile"       "fish"
## [5] "amphibian"     "insect"        "mollusc.et.al"
```

```
print(table(L.zoo))
```

```
## L.zoo
##        mammal          bird       reptile          fish     amphibian
##            41            20             5            13             4
##        insect mollusc.et.al
##             8            10
```

First, we need to commpute the distance matrix using the Euclidean metric.

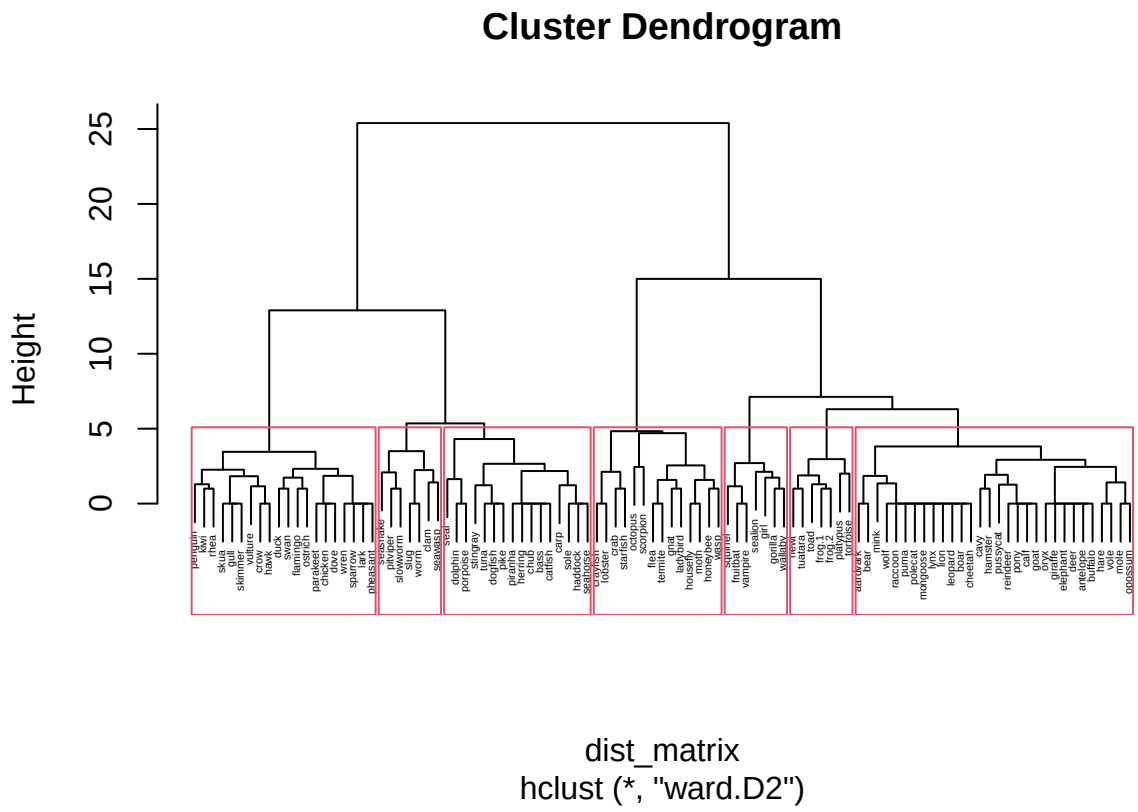
```
dist_matrix = dist(X.zoo, method = "euclidean")
```

We then perform hierarchical clustering using Ward.D2, average linkage, complete linkage and single linkage.

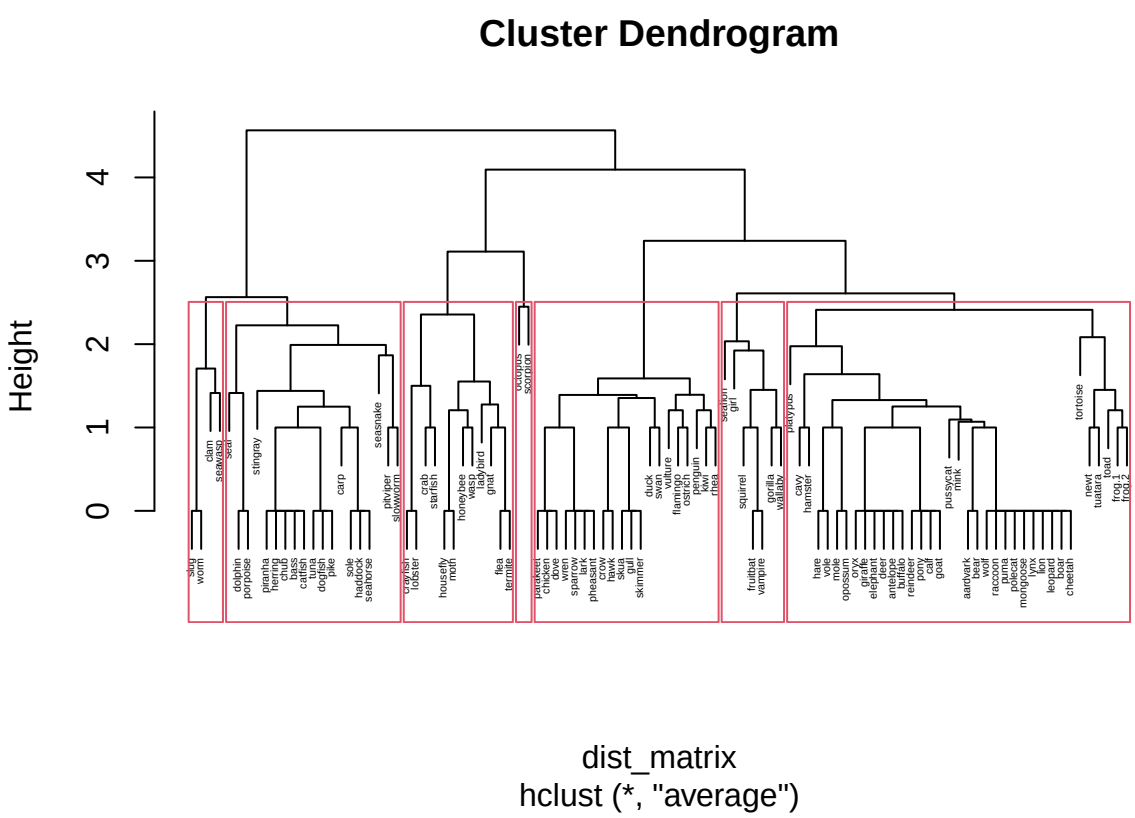
```
hclust.ward = hclust(dist_matrix, method = "ward.D2")
hclust.avg = hclust(dist_matrix, method = "average")
hclust.comp = hclust(dist_matrix, method = "complete")
hclust.sing = hclust(dist_matrix, method = "single")
```

Next, we plot the resulting trees a visualize the partition in seven clusters.

```
plot(hclust.ward, cex=0.3) # cex=0.3 reduce font size of leaf labels
rect.hclust(hclust.ward, k = 7)
```

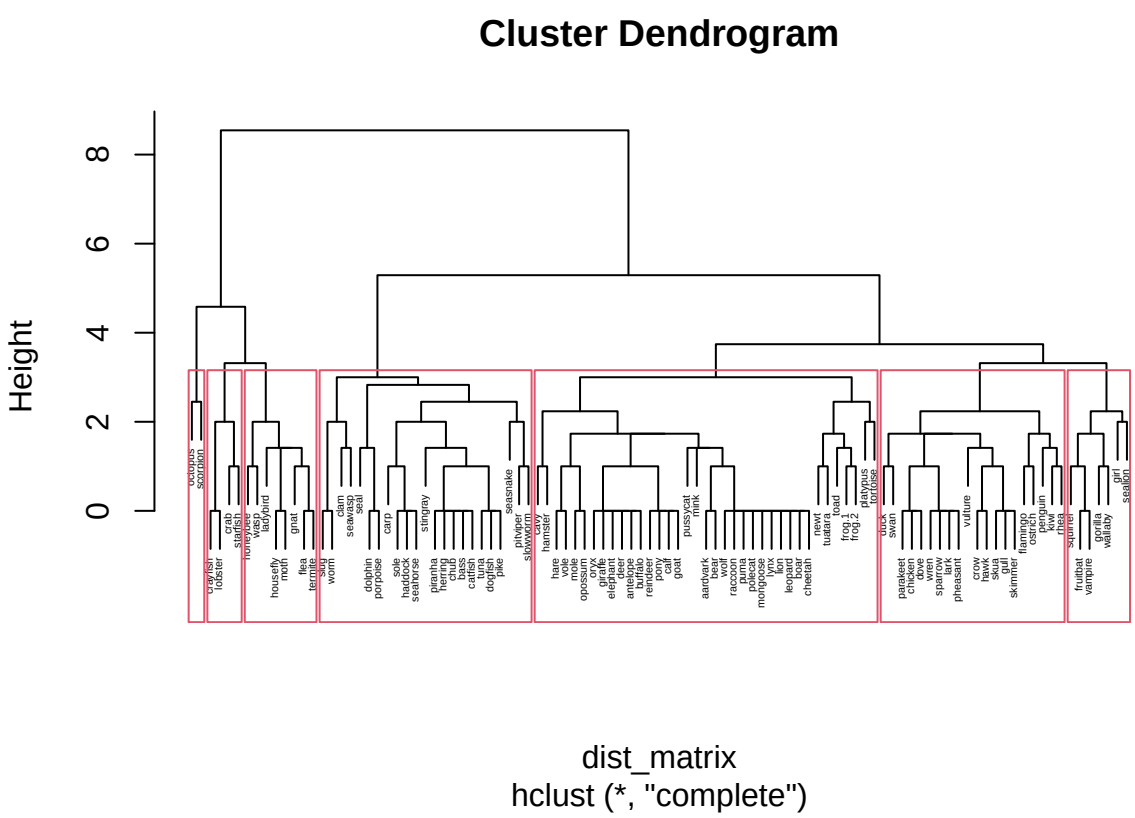



```
plot(hclust.avg, cex=0.3) # cex=0.3 reduce font size of leaf labels
rect.hclust(hclust.avg, k = 7)
```

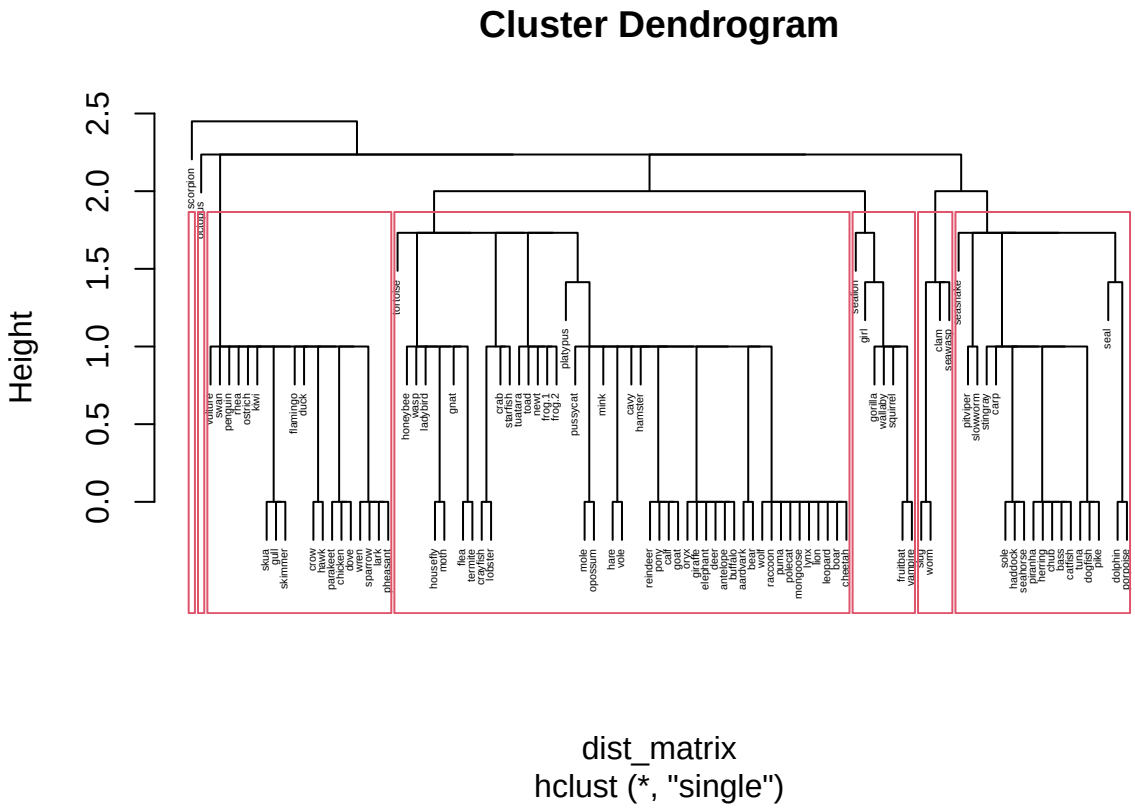


dist_matrix
hclust (*, "average")

```
plot(hclust.comp, cex=0.3) # cex=0.3 reduce font size of leaf labels
rect.hclust(hclust.comp, k = 7)
```




```
plot(hclust.sing, cex=0.3) # cex=0.3 reduce font size of leaf labels
rect.hclust(hclust.sing, k = 7)
```



Lastly, we will extract the allocation of samples to the seven groups for each method and compute the missclassification numbers compared to the ground truth.

```
groups.ward = cutree(hclust.ward, k = 7)
groups.avg = cutree(hclust.avg, k = 7)
groups.comp = cutree(hclust.comp, k = 7)
groups.sing = cutree(hclust.sing, k = 7)

print(table(L.zoo)) # the classes and the number of members in each class
```

```
## L.zoo
##        mammal          bird       reptile          fish     amphibian
##            41            20             5            13             4
##        insect mollusc.et.al
##             8            10
```

```
print(table(L.zoo, groups.ward))
```

```
##                groups.ward
## L.zoo            1  2  3  4  5  6  7
##   mammal        30  3  0  0  0  1  7
##   bird           0  0 20  0  0  0  0
##   reptile        0  0  0  3  0  2  0
##   fish           0 13  0  0  0  0  0
##   amphibian      0  0  0  0  0  4  0
```

```
##   insect         0  0  0  0  8  0  0
##   mollusc.et.al  0  0  0  4  6  0  0
```

```
print(table(L.zoo, groups.avg))
```

```
##                groups.avg
## L.zoo            1  2  3  4  5  6  7
##   mammal        31  3  0  0  0  7  0
##   bird           0  0 20  0  0  0  0
##   reptile        2  3  0  0  0  0  0
##   fish           0 13  0  0  0  0  0
##   amphibian      4  0  0  0  0  0  0
##   insect         0  0  0  0  8  0  0
##   mollusc.et.al  0  0  0  4  4  0  2
```

```
print(table(L.zoo, groups.comp))
```

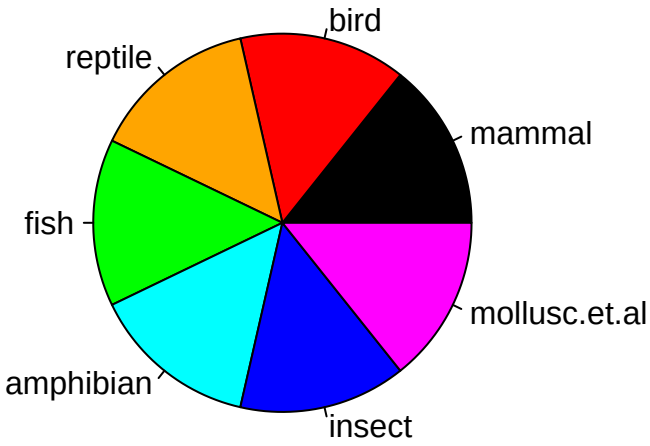
```
##                groups.comp
## L.zoo            1  2  3  4  5  6  7
##   mammal        31  3  0  0  0  7  0
##   bird           0  0 20  0  0  0  0
##   reptile        2  3  0  0  0  0  0
##   fish           0 13  0  0  0  0  0
##   amphibian      4  0  0  0  0  0  0
##   insect         0  0  0  0  8  0  0
##   mollusc.et.al  0  4  0  4  0  0  2
```

```
print(table(L.zoo, groups.sing))
```

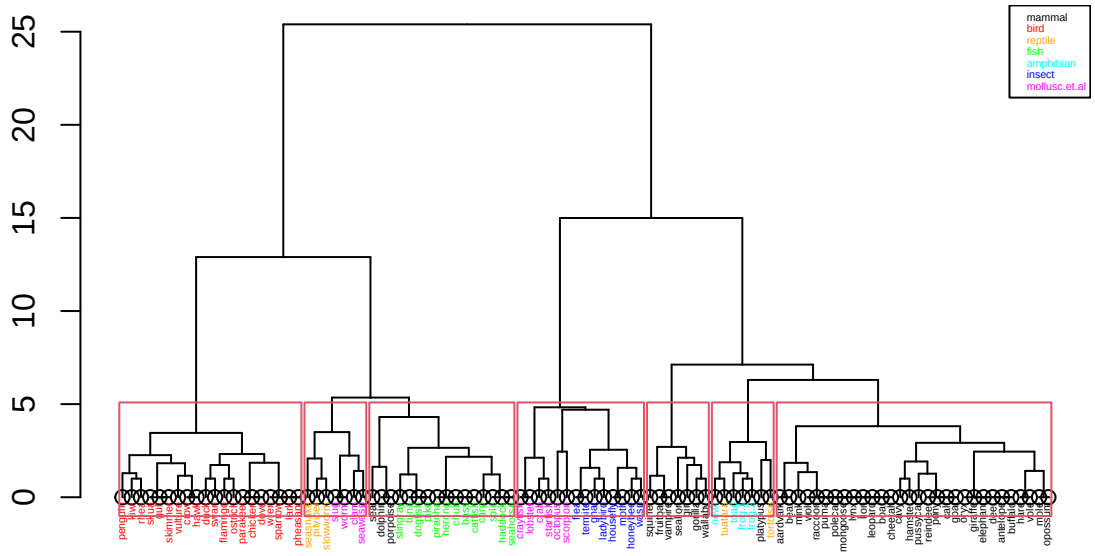
```
##                groups.sing
## L.zoo            1  2  3  4  5  6  7
##   mammal        31  3  0  0  7  0  0
##   bird           0  0 20  0  0  0  0
##   reptile        2  3  0  0  0  0  0
##   fish           0 13  0  0  0  0  0
##   amphibian      4  0  0  0  0  0  0
##   insect         8  0  0  0  0  0  0
##   mollusc.et.al  4  0  0  4  0  1  1
```

We can also plot the tree with coloured leaves to visualize the classes better.

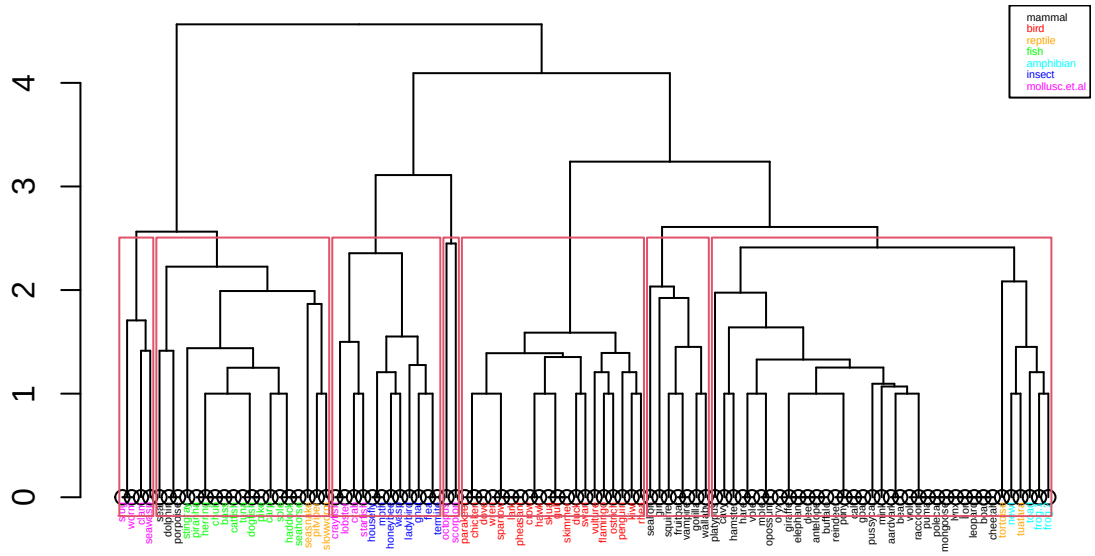
```
# define colors for the seven groups/levels in L.zoo
col.zoo = c("black", "red", "orange", "green", "cyan", "blue", "magenta")
pie(rep(1, 7), col = col.zoo, labels=levels(L.zoo))
```



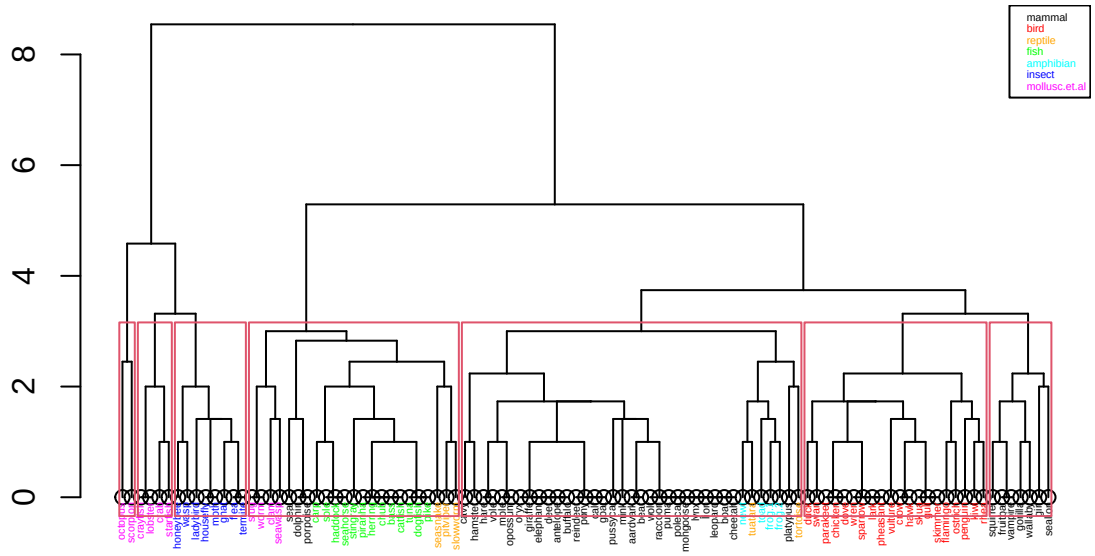

```
# name -> color
colMap = col.zoo[as.integer(L.zoo)]
names(colMap) = rownames(X.zoo)

# this function converts hc objects into dendograms, colors the leafs and then
# plots the tree along with a legend
plotColoredTree = function(hc, colMap, cex)
{
  colorLeafs = function(x)
  {
    if (is.leaf(x))
    {
      lbl = attr(x, "label") # label at leaf
      attr(x, "nodePar") = c(list(lab.col=colMap[lbl], lab.cex=cex)) # set color and font size
    }
    return(x)
  }
  hcd = dendrapply(as.dendrogram(hc), colorLeafs)
  plot(hcd, main=paste(hc$method, "+", hc$dist.method))
  legend("topright", legend=levels(L.zoo), text.col=col.zoo, cex=cex)
}

# plot the trees above with coloured leaves
plotColoredTree(hclust.ward, colMap, cex=0.3)
rect.hclust(hclust.ward, k=7)
```



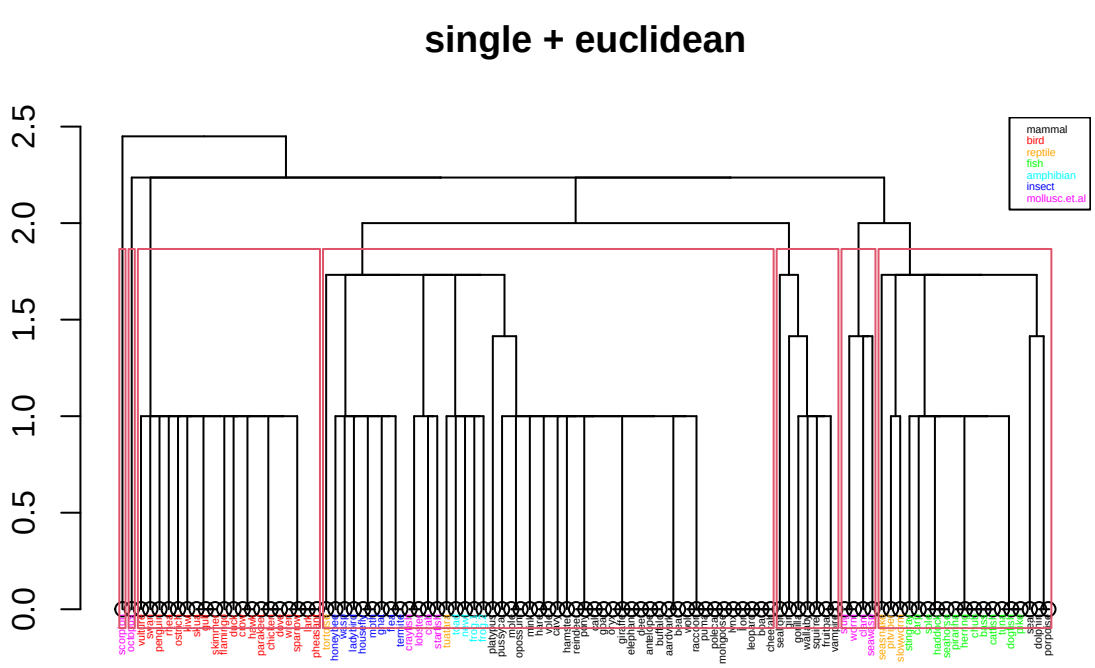

```
plotColoredTree(hclust.avg, colMap, cex=0.3)
rect.hclust(hclust.avg, k=7)
```




```
plotColoredTree(hclust.comp, colMap, cex=0.3)
rect.hclust(hclust.comp, k=7)
```



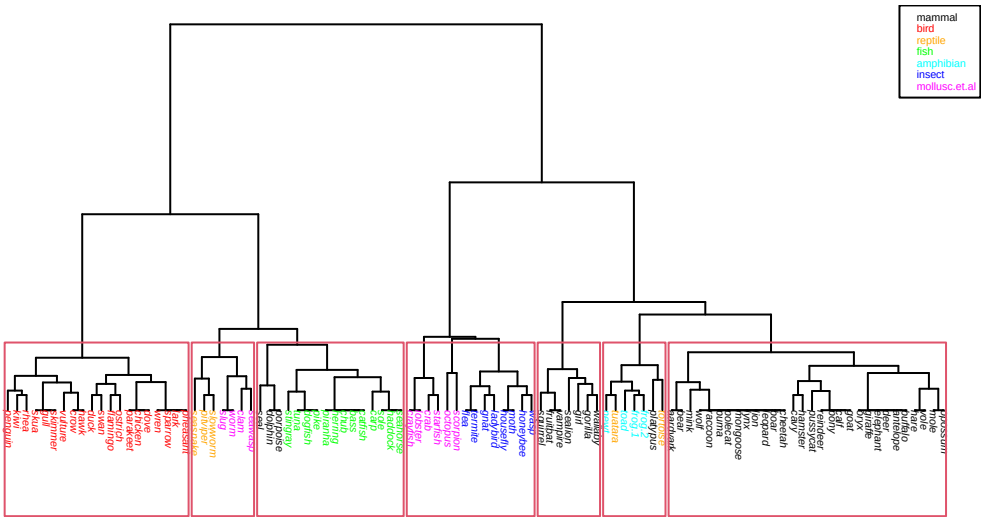

```
plotColoredTree(hclust.sing, colMap, cex=0.3)
rect.hclust(hclust.sing, k=7)
```



Another solution is to use the "ape" library to make the code more compact.

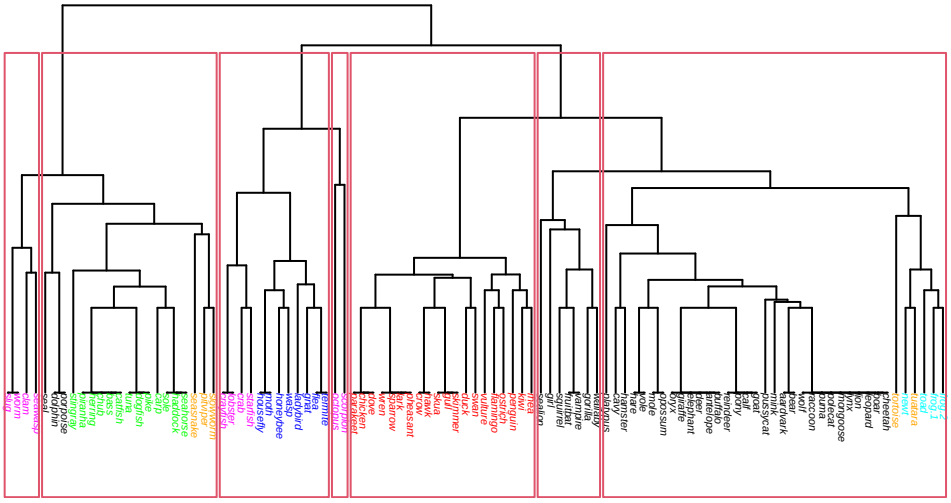
```
library("ape")

plot(as.phylo(hclust.ward), tip.color = colMap, cex = 0.3,
     direction = "downward", main=paste(hclust.ward$method, "+", hclust.ward$dist.method))
rect.hclust(hclust.ward, k=7)
legend("topright", legend=levels(L.zoo), text.col=col.zoo, cex=0.3)
```

ward.D2 + euclidean



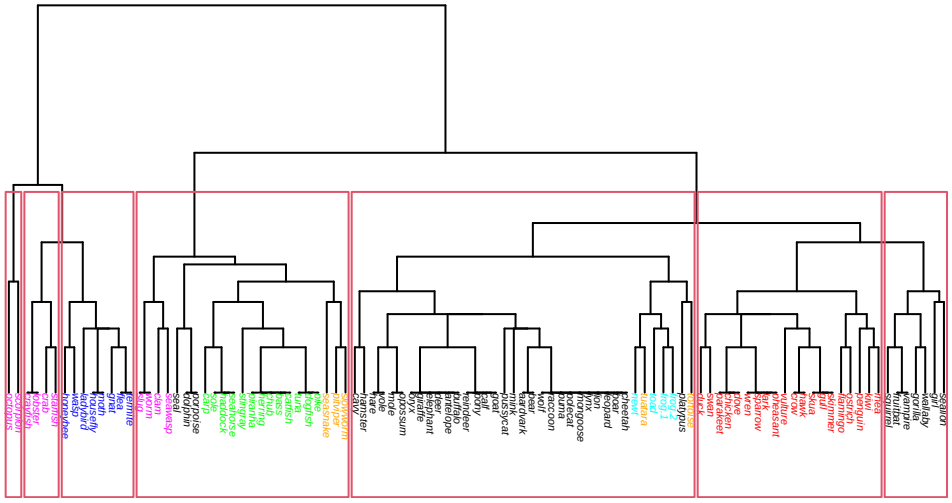

```
plot(as.phylo(hclust.avg), tip.color = colMap, cex = 0.3,
     direction = "downward", main=paste(hclust.avg$method, "+", hclust.avg$dist.method))
rect.hclust(hclust.avg, k=7)
```

average + euclidean



```
plot(as.phylo(hclust.comp), tip.color = colMap, cex = 0.3,
     direction = "downward", main=paste(hclust.comp$method, "+", hclust.comp$dist.method))
rect.hclust(hclust.comp, k=7)
```

complete + euclidean



```
plot(as.phylo(hclust.sing), tip.color = colMap, cex = 0.3,
     direction = "downward", main=paste(hclust.sing$method, "+", hclust.sing$dist.method))
rect.hclust(hclust.sing, k=7)
```

single + euclidean