R-CLAG: an unsupervised non hierarchical clustering algorithm handling biological data

Linda Dib[1,2], Raphaël Champeimont[1,2], Alessandra Carbone[1,2,*]

[1]Université Pierre et Marie Curie, UMR7238, 15, rue de l'Ecole de Médecine, 75006 Paris, France
[2]CNRS, UMR7238, Laboratoire de Génomique des Microorganismes, F-75006 Paris, France

August 19, 2013

Abstract

This package allows to use the CLAG clustering algorithm decribed in Dib and Carbone (2012).

Contents

1 Example with an artificial data set **2**

2 Globine coevolution matrix **2**

1 Example with an artificial data set

First, let's load the CLAG R package.

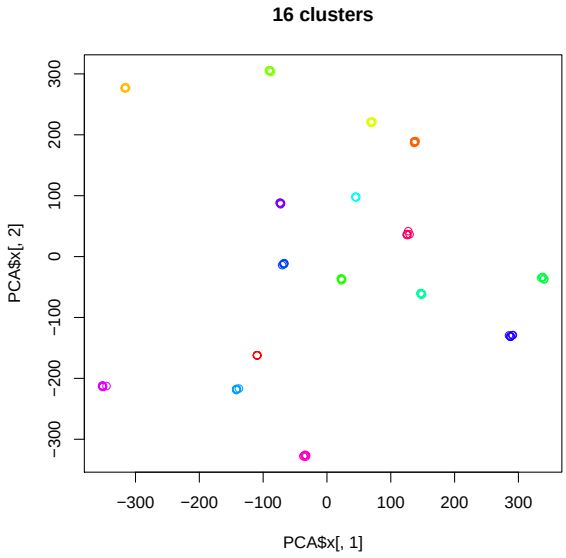
```
> library(CLAG)
```

Now, we load an example data set (provided with this package), then we run the CLAG algorithm on it.

```
> data(DIM128_subset, package="CLAG")
> RES <- CLAG.clust(DIM128_subset)
```

Now, we do a Principal Component Analysis to better visualize the data, and color the points according to the clusters found by CLAG (or leave black unclustered points).

```
> PCA <- prcomp(DIM128_subset)
> clusterColors <- c("black", rainbow(RES$ncluster))
> plot(PCA$x[,1], PCA$x[,2], col=clusterColors[RES$cluster+1],
+      main=paste(RES$nclusters, "clusters"))
```



As can be seen, the clusters are perfectly detected.

2 Globine coevolution matrix

This data set is a symmetric matrix of "coevolution scores" (some kind of correlation coefficients), we want to cluster its rows and colums.

Here we use analysis type 3, which means we want to use both environment score and **symmetric scores** (and not only environment score like by default).

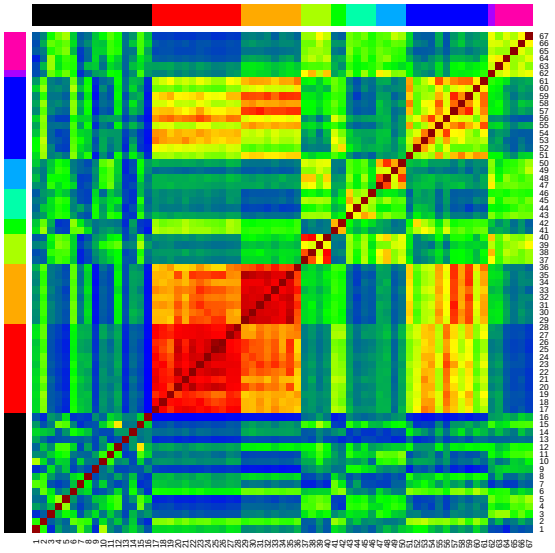
The symmetric score measure the position in the distribution of the original value in the matrix. The idea is that we are interesetd in clustering positions which have high scores between each other in the original matrix (and not simply rows which exhibit similar values for the same columns, like with environment score). This makes sense because the input matrix is a correlation-like matrix.

We load the data and run the cluster analysis, with the parameters proposed in Dib and Carbone (2012).

```
> data(GLOBINE, package="CLAG")
> M <- GLOBINE$M
> RES <- CLAG.clust(M, delta=0.2, threshold=0.5, analysisType=3)
```

Now reorder the rows and columns to group them by cluster. Then plot the matrix with bars on left and right where the color indicates the cluster (black is used for unclusted elements).

```
> o <- order(RES$cluster)
> M2 <- M[o,o]
> clusterColors <- c("black", rainbow(RES$nclusters))[RES$cluster[o]+1]
> colorScale <- colorRampPalette(c("blue", "green","yellow","red","darkred"))(1000)
> heatmap(M2, symm=TRUE, Colv=NA, Rowv=NA, scale="none", col=colorScale,
+         ColSideColors=clusterColors, RowSideColors=clusterColors)
```



Notice that groups of elements which have a high correlation between each other are clustered together, and that elements which have low correlation with every other are left unclustered.

References

Dib, L. and Carbone, A. (2012). CLAG: an unsupervised non hierarchical clustering algorithm handling biological data. *BMC Bioinformatics*, 13(1):194.