ex7: Clustering

演習問題1

データセットSato_A_thaliana-P_syringae_arvRpt2_6h_expRatio_small.txt(61遺伝子x8遺伝子型）を使って、ユークリッド距離を使った場合とコサイン係数を使った距離でクラスタリングした時の違いを調べなさい（クラスタリング結果の違いの可視化はdendextendライブラリーを使うのが便利である）。このデータはシロイヌナズナ変異体 にバクテリアを感染させた際の発現プロファイルを取り、野生型との比（log2)をとったものである。なお、コサイン係数での距離、クラスタリングは下記 のカスタム関数を用いてよい。また、heatmapおよびheatmap.2(gplotsライブラリー）では引数に Rowv=as.dendrogram("行のクラスタリング結果"), Colv=as.dendrogram("列のクラスタリング結果")と指定することで任意のクラスタリング結果でヒートマップを描くことができる。

```
library(colorspace)
library(dendextend)
library(dendextendRcpp)
library(gplots)

inputMatrix   <- read.delim("~/data/MS/Sato_A_thaliana-P_syringae_arvRpt2_6h_expRatio_small.txt",
                 header=TRUE, row.name=1)
heatmapColors <- colorpanel(10, low="blue", mid="white", high="orange")

heatmap.2(as.matrix(inputMatrix),
          scale="none",           # 発現量比のスケーリング無し
          trace="none",           # heatmap.2デフォルトのトレースをキャンセル
          # ヒートマップのマス目設定
          sepcolor="black", colsep=0:ncol(inputMatrix), rowsep=0:nrow(inputMatrix), sepwidth=c(0.01, 0.01),
          density.info="none",    # ヒストグラム
          col=heatmapColors,
          cexRow=(0.2 + 1/log10(nrow(inputMatrix)))/3*2,
          RowSideColors=ifelse(rownames(inputMatrix)=="At2g14610", "magenta", "grey")
)

# コサイン係数（ベクトルの角度）でクラスタリング
# コサイン係数は関数が無いので自作する
cosine.coef <- function(x,y) {
    a <- sum(na.omit(x * y)) / sqrt( sum(na.omit(x)^2) * sum(na.omit(y)^2) )
    return(a)
}

# making a distance table between columns using uncentered Pearson correlation
cosine.table <- function(x) {
    numberOfPoints <- ncol(x)
    columnNames    <- colnames(x)
    distanceTable  <- matrix(data = NA, nrow = numberOfPoints, ncol = numberOfPoints,
                             dimnames = list( columnNames, columnNames )
                            )

    for ( i in 1:(numberOfPoints-1) ) {
        for ( j in (i+1):numberOfPoints ) {
            v1 <- x[ , i]
        v2 <- x[ , j]
        d  <- 1 - cosine.coef(v1, v2)
        distanceTable[i, j] <- d
        distanceTable[j, i] <- d
    }
    }

    for ( i in 1:numberOfPoints ) { distanceTable[i, i] <- 1 } # fill the diagonal
        return(distanceTable)
}

# コサイン係数の距離行列
# cosineDistanceTable <- as.dist(cosine.table(as.matrix((dat1))))

# 行のクラスタリング
```

```
rowClusters <- hclust(as.dist(cosine.table(as.matrix((t(inputMatrix))))))
# 列のクラスタリング
colClusters <- hclust(as.dist(cosine.table(as.matrix((inputMatrix)))))

heatmap.2(as.matrix(inputMatrix), Rowv=as.dendrogram(rowClusters), Colv=as.dendrogram(colClusters),
  scale="none", trace="none", sepcolor="black", colsep=0:ncol(inputMatrix), rowsep=0:nrow(inputMatrix),
  sepwidth=c(0.01, 0.01), density.info="none", col=heatmapColors,
  cexRow=(0.2 + 1/log10(nrow(inputMatrix)))/3*2,
  RowSideColors=ifelse(rownames(inputMatrix)=="At2g14610", "magenta", "grey"))


rowClusters1 <- as.dendrogram(hclust(as.dist(dist(as.matrix((inputMatrix))))))
rowClusters2 <- as.dendrogram(hclust(as.dist(cosine.table(as.matrix(t(inputMatrix))))))
rowDendrogramList <- dendlist(rowClusters1, rowClusters2)
png("tanglegram_row.png", width=480*4, height=480*2, res=200)
tanglegram(rowDendrogramList, common_subtrees_color_branches = TRUE, columns_width= c(10,3,10),
  lab.cex=0.6, lwd=2, main_left="Euclidian", main_right="Cosine")
dev.off()


colClusters1 <- as.dendrogram(hclust(as.dist(dist(as.matrix(t(inputMatrix))))))
colClusters2 <- as.dendrogram(hclust(as.dist(cosine.table(as.matrix((inputMatrix))))))
columnDendrogramList <- dendlist(colClusters1, colClusters2)
tanglegram(columnDendrogramList, common_subtrees_color_branches = TRUE)
```