Tutorial 2

The "ClueR" R package contains a time-course phosphoproteomics dataset "hES". Each column of in hES data is a time point and each row is a phosphorylation sites. We will perform clustering analysis on this dataset.

(1) Install "ClueR" R package and its dependent packages. Find out how to use it by typing "?runClue".

(2) Once you have installed the package load the hES dataset as follows:

```
data(hES)
```

Find out the dimension of the hES dataset.

(3) Create hierarchical clustering with respect to times (i.e. cluster the columns). How does time points cluster with each other? Does it make sense?

(4) Install package "e1071" and apply c-means clustering to partition the data in to 9 groups (c=9) with respect to phosphorylation sites (i.e. partition rows into c groups). Firstly, standardise the data to be unit free.

```
standardize <- function(mat) {
    means <- apply(mat, 1, mean)
    stds <- apply(mat, 1, sd)
    tmp <- sweep(mat, 1, means, FUN="-")
    mat.stand <- sweep(tmp, 1, stds, FUN="/")
    return(mat.stand)
}

hES.scaled <- standardize(hES)
```

Once the data is standardised the data to be unit free, perform clustering.

```
library(e1071)

fc <- cmeans(hES.scaled, centers=9)
```

Visualise the clustering results using ClueR package function "fuzzPlot" as follows:

```
fuzzPlot(hES.scaled, fc, mfrow = c(3, 3))
```

(5) Is k=9 the best choice of k? Apply Dunn index to validate k-means clustering using different k values. Which K gives best clustering results according to Dunn index? Does it differ if we use other validation index such as Connectivity or APN?