Assignment #3
(data analysis)

The data we consider are from:
Gasch A.P., Spellman P.T., Kao C.M., Carmel-Harel O., Eisen M.B., Storz G., Botstein D., Brown P.O. (2001), Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes, *Molecular Biology of the Cell* **11**, 4241-4257.

Expression is recorded for N=6152 known and putative yeast genes, on over 140 conditions. We concentrate on a T=8 time course following a heat shock from 25 to 37C. The time points correspond to minute 5, 10, 15, 20, 30,40, 60, 80 after the shock. The values are log-ratios to a baseline obtained pooling equal amounts of all experimental samples. The profiles of 2509 genes (40.78% of the total) have missing values.

In the file Yeast_shock.XLS you will find a 6152 by 8 data matrix (plus gene names and short descriptions), in which missing values were imputed through a mixture model fit. THUS, YOU DO NOT HAVE TO WORRY ABOUT MISSING VALUE IMPUTATION FOR THIS ASSIGNMENT.

1. Decide whether to apply centering and standardization by row (gene) and/or by column (time point). Give an argument for your choice (possibly related to the outcomes of the analyses requested below).
2. Extract the principal components of the data, produce plots, and comment on the results.
3. Chose one of the following
    - Design and perform a (small) re-sampling analysis to assess the sample variability/stability of the eigenvalues and eigenvectors in PCA. Comment on the results.
    - Perform a multidimensional scaling, and compare the results to those of the principal component analysis.
4. Cluster the data. You can chose between k-means and hierarchical clustering (with a given distance and link selection), and between clustering the data in the original 8 dimensions, or within a reduced representation obtained through principal components or multidimensional scaling. Again, give an argument for your choices. Produce plots, and comment on the results. As for the choice of number of cluster, you can, if you want, just provide a heuristic argument based on similarity along the dendrogram, or within cluster sum of squares, but you are encouraged to perform a (small) re-sampling analysis along the lines proposed in Ben-Hur *et al*. (you can use any partition similarity measure you want; matching index, their correlation index, etc.).

Note: Whenever you are asked to comment, you should try to make “biological sense” of the results. However, a detailed analysis in terms of identification of individual genes in clusters, with their functional and/or regulatory relationships, is NOT required for this assignment.