

# Data analysis assignment: Cluster 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.**

In this study, 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 normalized log-ratios to a baseline obtained pooling equal amounts of all experimental samples.

In the original data, the profiles of 2509 genes (40.78% of the total) have missing values. However, in the file **yeast\_shock.txt** you will find a 6152 by 8 data matrix, plus gene identifiers (short descriptions are also available in **yeast\_shock.xls**), 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. Pre-processing

- a. Produce histograms and normal q-q plots for each time point (i.e. data column), to ascertain the effectiveness of the normalization that was applied to these data (see Yang et al. 2001): Do the histograms look centered at 0, bell shaped and fairly “regular”? Do they present very different spreads?
- b. Decide whether to apply centering and standardization by row (gene) and/or by column (time point) prior to clustering. Give an argument for your choice.
- c. Decide whether to “filter out” some of the genes prior to clustering. Again, give an argument for your choice. Hint: you could filter based on the variability presented by each gene across the 8 time points (a statistic to use could be the gene coefficient of variation, i.e. the sd divided by the absolute value of the gene mean across the 8 time points); remember that a filtering of this type needs to be performed prior to row standardization.

## 2. Clustering

- a. Chose a clustering algorithm. i.e. K-means, or hierarchical clustering with a given distance and link function. You can, if you want, consider more sophisticated algorithms, but you are not required to. Give an argument for your choice.
- b. Chose between clustering the data in the original 8 dimensions, or within a low-dimensional representation obtained through principal components. Give an argument for your choice.
- c. Chose the number of clusters (**this is the part of the assignment that will require the most work, and likely some coding**). Using Dudoit and Friedlyand (2002) as a reference, select an internal index (produce the corresponding plot on  $k = \#$  of clusters; describe and implement the choice of  $k$ ). Alternatively, implement a perturbation/re-sampling analysis based on an external index, along the lines described in Ben-Hur et al (2002) and Dudoit and Friedlyand (2002) – you can be creative if you wish, and a perturbation/re-sampling study need NOT be large for this assignment.
- d. Produce tables and plots summarizing the clustering output, and comment on the results – you should try to make “biological sense”, but a detailed analysis of individual genes in clusters, with their functional and/or regulatory relationships, is NOT required for this assignment.