Clustering Homework, Bio 311, Spring 2017

Instructions: Carry out the analyses described below in R, creating a report in the form of an R Markdown document. Submit the final R markdown document (the `.Rmd` file NOT the HTML output) via Sakai.

1. Download the file `gasch1k.csv` from Sakai

    - This is a subset of the data described in the following paper: Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown, PO. 2000. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell*. 11(12):4241-57.

    - I have pre-filtered the data set, removing any genes for which there were too many missing values and reducing the data set to the most variable genes.

2. Exploring the data

    a) How many genes are represented in the data set?

    b) How many samples are in the data set?

    c) EXTRA CREDIT: The row names of the data correspond to the experimental conditions (perturbations) that were carried out. Each row name is of the form "Heat.Shock.05.minutes....". If you treat the first "word" of the name (e.g. "Heat") as indicating the class of the experiment, how many classes of perturbation are there? You might find the `stringr` library (http://stringr.tidyverse.org , installed by default on VM Manage) useful to answer this question.

3. Visualizing a subset of the data

    a) Create a subset of the data corresponding to Nitrogen Depletion time series experiments (row names that sart with "Nitrogen.Depletion....")

    b) Create a new data frame with an additional column corresponding to the "time" associated with each nitrogen depletion measurement (the `mutate` function in dplyr might be useful)

    c) For the first 10 genes in the data set, create time series plots, showing how the expression of each gene changes over time in the nitrogen depletion experiments.

4. Hierarchical clustering

    a) Carry out hierarchical clustering on the genes in the data set using "complete linkage" clustering using (1 - Pearson correlation) as the distance metric

- Generate trees and heatmaps to illustrate the clustering you generated

- Explore how cutting the resulting dendrogram are different heights changes the number and size of the resulting clusters.

- In addition to automatic cutting based on tree height, it's perfectly legitimate to consider using visual inspection or other "manual criteria" to further subdivide clusters. Note that if you generate sub-trees with the `cut` function, you can iteratively apply cuts to sub-trees to explore sub-structure.

- Create a correlation heat-map, with the rows and variables of the matrix sorted by cluster assignment, to help evaluate your clustering.

- Based on your inspection and analysis of the data, how many major clusters (>20 genes) do you think there are? Explain your reasoning for arriving at this number.

b) Repeat the hierarchical clustering, but with "single linkage" as the clustering method instead

- Generate trees, heatmaps, and correlation heatmaps to illustrate the single clustering you generated

- As before, vary the height at which you cut the dendrogram. How many many clusters do you think there are based on the single linkage dendrogram?

- Which clustering do you think is more useful in terms of exploring and understanding the data – the complete or single linkage? Why?

5. K-medoids clustering

   a) Carry out K-medoids clustering on the genes, using the number of major clusters you determined via hierarchical clustering.

   b) Re-run the K-medoids clustering, varying the number of clusters. How do the results change your interpretations of the data?

6. Gene ontology

   a) For the three largest clusters you generated via K-means clustering, carry out a GO term enrichment analysis based on the "Biological Process" ontology, using the g:Profile web tool at http://biit.cs.ut.ee/gprofiler/

   b) What terms did you find to be enriched in your clusters? For more compact representation of the ontology output from g:Profiler try setting the "Hierarchical filtering" setting to "Best per parent (moderate").

7. Clustering rows (conditions)

   Rather than clustering the genes (columns of the data frame), carry out hierarchical clustering on the experimental conditions (rows of the data set). Does the clustering of experimental conditions make sense in terms of the types of perturbations that were used? What perturbations cluster most closely with the Nitrogen Depletion experiments?