Homework 6: Due Dec 8th 2023

1. Use the data set “120217\_cellCycleData.txt” and perform KNN imputation. Use only the alpha-factor synchronized data (the 6th to 23rd columns) in the data set.
   * 1. Project the 800 genes onto two-dimensional space using PCA and MDS respectively. Label the genes of different functional annotation (“G1”, “S”, “S/G2”, “G2/M”, “M/G1”) with differential colors (col=1~5) and text symbols (1~5).
     2. Perform hierarchical clustering (K=5) and K-means (K=5). Compare clustering result from each algorithm to the truth given in the last column of the data (“G1”, “S”, “S/G2”, “G2/M”, “M/G1”). Calculate adjusted Rand indexes (“adjustedRandIndex” in “mclust” package) for the comparisons. Which is better? Summarize what you observe.
     3. Plot the heatmap of cluster patterns from each method.
2. In this problem we are going to practice pathway analysis. Let’s re-use the DE analysis results from in HW4 1(e). For pathway database, load in the GO gene set from file “c5.all.v2.5.symbols.gmt”. In order to convert the identifier in Golub\_Merge to gene symbol, you need to download package “hu6800” from Bioconductor. Use function “hu6800SYMBOL” to achieve this goal.
   * 1. How many gene sets are available from “c5.all.v2.5.symbols.gmt”? Overlap genes available in this file and Golub\_Merge which is used in HW4. How many genes are overlapped? Update the gene sets by dropping genes that do not appear in the Golub\_Merge.
     2. Filter out gene sets that contain less than 5 genes or more than 200 genes. How many gene sets are left?
     3. Perform Fisher’s exact test for each gene set on the DE genes derived in HW4 1(e). Pay attention to the background gene that you use. How many pathways are significant? List the top 10 pathways. Comment on the result.