



# Northeastern University

## College of Science

### Module 2 Homework

#### Problem 1. (30 points)

Computations on gene means of the Golub data set.

- (a) Compute the mean expression values for every gene among “ALL” patients.
- (b) Compute the mean expression values for every gene among “AML” patients.
- (c) Give the biological names of the three genes with the largest mean expression value among “ALL” patients.
- (d) Give the biological names of the three genes with the largest mean expression value among “AML” patients.

Submit R commands that does (a)-(d). And answer directly part (c) and (d)



# Northeastern University

## College of Science

### Problem 2. (30 points)

More work on the Golub data set.

- (a) Save the expression values of the first five genes (in the first five rows) for the AML patients in a csv file "AML5.csv".
- (b) Save the expression values of the first five genes for the ALL patients in a plain text file "ALL5.txt".
- (c) Compute the standard deviation of the expression values on the first patient, of the 100<sup>th</sup> to 200<sup>th</sup> genes (total 101 genes).
- (d) Compute the standard deviation of the expression values of every gene, across all patients. Find the number of genes with standard deviation greater than 1.
- (e) Do a scatter plot of the 101th gene expressions against the 102th gene expressions, label the x-axis and the y-axis with the genes' biological names using xlab= and ylab= control options.

Submit R commands that does (a)-(e). And the outputs (files for parts (a), (b), numerical answer for part (c) and (d), the figure file for part (e)).



# Northeastern University

## College of Science

### Problem 3. (20 points)

Work with the ALL data set. Load the ALL data from the ALL library and use `str` and `openVignette()` for a further orientation.

- (a) Use `exprs(ALL[,ALL$BT=="B1"])` to extract the gene expressions from the patients in disease stage B1. Produce one histogram of these gene expressions in the this matrix.
- (b) Compute the mean gene expressions for every gene over these B1 patients.
- (c) Give the gene identifiers of the three genes with the largest mean.

Submit R commands that does (a)-(c), and answer part (c) directly.



# Northeastern University

## College of Science

### Problem 4. (20 points)

We work with the “trees” data set that comes with R.

- (a) Find the type of the trees data object.
- (b) Produce a figure with two overlaid scatterplots: Height versus Girth, Volume versus Girth(The Girth is on the x-axis). Do the Height plot with blue “+” symbols, and do the Volume plot with red “o” symbols. You need to learn to set the ylim= control option so that all points from the two plots can all show up on the merged figure.

Hint: you should use plot() then points() to create the overlaid two scatterplots.