BIOLOGY 664: Integrated Bioinformatics Using R for Both Wet and Dry Scientists

Problem Set 1: Creating and manipulating data.frames

Due: At the beginning of class Thursday, February 26th.

Turn-in: Both a printed copy in class and a softcopy into Blackboard.

Part 1

Merge the most relevant data found in the 3 tables (golub.gnames, golub, and golub.cl) that make-up the golub data in the library(multtest) into one data.frame with the following properties:

Name: golub.df
Dimensions: patient rows and named gene columns, and an additional named column for the cancer classifications
Column Names: use the gene name (column 2) from golub.gnames and “classification”
Classification Column: use a factor column in golub.df that uses "ALL" and "AML" as the classifications

Part 2

Answer the Chapter 1 Exercises 3 through 5 using your new golub.df data.frame – except for exercises 3c and 5d. Try not to cheat by reformulating the answers in the book – unless you get really stuck.

Notes and Hints:

1. Don't use the gene index or gene ID (columns 1 and 3 in golub.gnames) in your new data.frame. Just have named gene columns and one extra named column for the cancer classification.
2. Use t() to transpose a matrix
3. Use [!names(golub.df) %in% c("column1", “column2”, …)] to remove columns from a data.frame
4. You can work together, but all your written work (including R code) must be your own.