**Exercise 4**

Complete this exercises as a group and save your answers in a well-commented R script. Be sure to discuss each question amongst all members of the group, and also feel free to explore multiple solutions to the same problem. I expect you will need to consult class notes and R help files. You are welcome to use any printed or web-based resources, as well. Groups will be randomly selected to present their answers in the next class meeting.

1. You conducted an experiment over the summer where you measured the expression of a stress response gene in Daphnia under three temperature levels (T1, T2, T3). Your undergraduate research assistant is now back at their home institution, but left you the data file “ThankYouREU.txt”. Unfortunately, the research assistant wasn’t very careful in labeling the treatments and the labels are a bit of a mess. The assistant assures you that data labeled “T4” comes from T3 experimental units. Use your string manipulation skills to cleanup this data table.
2. You want to find the number and length of putative genes in the dna sequence contained in “dna.txt”. You know that genes start with “ATG” and end with “TAG”, “TAA”, or “TGA”. Create a script that uses regular expressions to find the number of genes and length of each gene in “dna.txt”.
3. You hypothesize that population density in Indiana Counties is correlated with the percent of land covered by lakes, rivers, and streams. You collect percent surface water land cover data from a GIS database (“IndianaCountyPercentWater.txt”) and you collect population density data from the Census Bureau (“IndianaCountyPopDensity.txt”). Unfortunately, the two data sources use different formats for county names. Reformat one set of names and merge the data into a single data frame.
4. You are provided files containing daily precipitation for the year 2010 from each of 20 major cities in the U.S. Create a script that generates a summary table with minimum daily, maximum daily, and total precipitation for each city.