CSC / BIO 310
Bioinformatics
Instructor: Dr. Laurie J. Heyer
Assignment #3
Due Thursday, Jan 31

Instructions: Create a Word document to answer the questions below, and email the Word document to me. You may not consult with anyone outside of your team, other than me. To do this assignment, you will need to modify the files regex.pl or regex_whole_genome.pl with the appropriate regular expressions described in chapters 4 and 12.

1.
    a. Explain what an EcoRI site is, and how these sites can be used by biologists.
    b. Use a regular expression to find all EcoRI sites in the E. coli genome. Paste the line of your Perl script containing this regex into your Word document.
    c. List the number of EcoRI sites that were found in the genome.

2.
    a. Modify either regex.pl or regex_whole_genome.pl to read the file TOM2_GeneList_noblank.txt rather than the E. coli data.
    b. Write a regular expression that finds lines that start with "blank". Paste the line of your Perl script containing this regex into your Word document.
    c. List the number of lines in TOM2_GeneList_noblank.txt that start with "blank", and the number of lines that do not start with "blank".

3.
    a. Do a web search to find a disease associated with dinucleotide repeats.
    b. Use regular expressions to find all occurrences of dinucleotide repeats in *E. coli*. List the number of occurrences for each observed number of repeats.

4. Use regular expressions to find all occurrences of strings of dinucleotides in which each dinucleotide consists of a repeated base. TTGGTTTT and AATTGGAACC are examples of this motif, of length 6 and 10, respectively. List the number of non-overlapping occurrences of this motif of each length greater than or equal to 8.

5. Explain the similarities and differences in each of the following regular expressions:
    a. ((.)(.)(.)\4\3\2)
    b. ((.)(.)(.)\4\3\2)\1
    c. ((.)(.)(.)\4\3\2)\1*
    d. ((.)(.)(.)\4\3\2)+