

CSC / BIO 310
Bioinformatics
Instructor: Dr. Laurie J. Heyer
Assignment #2
Due Thursday, Jan 24

Write comments in your file to indicate the solutions to each of the following problems. You should also include the number of each problem in the print statement.

Write a comment at the top of the program containing the five lines at the top of this assignment page, in addition to the names of people in your programming team.

You may not consult with anyone outside of your programming team, other than me.

In addition to the accuracy of your solutions, you will be graded on the readability of your code and your output, and the use of good programming practices as discussed in the text and in class.

To do this assignment, you will need to modify the file `regex.pl` with the appropriate regular expressions, as described on pages 53-55 and illustrated on pages 58-69 of the textbook.

1. How many words in `ecoli.txt` start with each possible dinucleotide? Write the regular expression for finding each list in a Word document. How many 7-mers should start with each dinucleotide if all possible 7-mers are present in the file?
2. Describe a procedure for finding all missing 7-mers in a list.
3. Apply your procedure to find the missing 7-mer in the `ecoli.txt` file.
4. Find all words in `ecoli.txt` that contain AGTCAT. Capture your output and include it in your Word document. Count the number of lines in your output file, and report this number.
5. Find all words in `ecoli.txt` that contain a start codon preceded by a C or G, and followed by an A or T. Capture your output and include it in your Word document. Count the number of lines in your output file, and report this number.