CS241 – Data Organization
Fall 2020

Programming Assignment #5: DNA

Due by 11:59 PM on September 27

In this assignment, you will write a simple C program that includes processing input and using control structures

Log into the host **linux.unm.edu.** Create a directory for this assignment as follows:

$ cd cs241 ; mkdir prog05

$ chmod 700 prog05 ; cd prog05

DNA is the basis of life here on earth. Every living creature is described by their DNA, which is a very, very long sequence of the bases A, C, G and T. Bioinformatics is a new discipline that uses the power of computers to analyze biological data, to discover new insights about the mysteries of life. For this assignment, you will complete a very simple bioinformatics experiment.

The input for your program will be a text file consisting of a sequence of DNA bases. Each base can appear as either a upper-case or as a lowercase letter (i.e., either 'A' or 'a' is legal). Your program must read the file, character by character, counting the total number of occurrences of each type of base. Your program will then output, for each base, what percentage of the DNA sequence consists of that base. For example, if the input is the sequence: **ACGGGTCGAC** then the output will be **exactly** these five lines of text:

The DNA sequence has 10 bases

20.00% of the bases are A

30.00% of the bases are C

40.00% of the bases are G

10.00% of the bases are T

The output must display the percentages with **exactly** three places to the left of the decimal point (to allow for 100%). The output must display the percentages with **exactly** two places to the right of the decimal point. The percent signs in the output must all be vertically aligned.

If your program successfully completes its task, then the program should return an exit status of 0 to the operating system. The input data file may contain any arbitrary "white space" (tabs, blank spaces, newlines) intermixed with the characters representing the DNA bases. These chars should not be counted. If the input data contains any other symbols, then your program should abort processing, and return an exit status of 1 to the operating system.

Since there are four kinds of bases, this program is a very natural place to use a switch statement. Your program **must** incorporate a **switch**.

Your program **must** read from the standard input (stdin), and not from any other file.

You can copy the files provided for this assignment using the following command (do not forget to include the dot symbol at the end of this command):

```
$ cp   ~soraya/cs241_Soraya/programs/ DNA/handouts/* .
```

I have provide four input files and the corresponding correct output files for three of them; one of the input files has other symbols. If you develop a program called foo.c for this assignment, then you can test your program using these test data files as follows:

```
$  gcc  foo.c

$ ./a.out   <   dnaInput_1  > myOutput

$  diff   myOutput  outputFile_1
```

The output of your program should be **char-by-char identical** to the given outputFile files, for those input files with correct input. The **diff** command will display any places where the two argument files are different. If the **diff** command is silent, then you know the two files are the same. We will also run your program with other input to make sure it behaves according to the specifications given in this statement.

If at any time when you are running your program, and you are "stuck" (the program has fallen into an infinite loop, or is unresponsive) you can "bail out" and halt the execution of the program using ^C (i.e., hold Control and C simultaneously).

You should be sure to include **THE NAME OF THE AUTHOR OF THE PROGRAM** in a comment at the top of your source code file, for this and all other assignments in this course.

Your source code must use proper **style**, that is variables should be well named (name is not too short, not too long, and is meaningful), and bodies of loops, if's, etc.. should be properly indented. Refer to the coding style file for this class, published on Learn under Coding Standards in a file named: cs241_codingStandards2020.pdf

Create a .c file for this assignment and name it using your last name and the initial of your first name, like this: **lastName_initialFirstName_DNA.c** Submit this file for grading to Learn in the place of this assignment.