CPSC 350: Data Structures
Fall 2019
Programming Assignment 1: C++ Review
Due: Sept. 14th, 2019. 11:59pm

The Assignment

For this assignment, you will build a simple analysis program that will compute basic statistics for a list of DNA strings. Your program should work as follows:

-The program will accept as a command line argument the name of a text file that will contain an arbitrary list of DNA strings. (ie. ~/assign1/filename.txt) DNA strings consist of a sequence of nucleotides (A,C,T, or G). There will be 1 string per line of the file. No guarantees on capitalization.

-The program will then compute the sum, mean, variance, and standard deviation of the length of the DNA strings in the list. It will also compute the relative probability of each nucleotide (A,C,T, or G), as well as the probability of each nucleotide bigram (AA, AC, AT, AG, CA, CC, CT, CG, etc) across the entire collection.

-The program will output the labeled results to a file called *yourname*.out. At the top of the file, output your name, student id, etc.

-After printing the summary statistics to *yourname.out*, you will generate 1000 DNA strings whose lengths follow a Gaussian distribution with the same mean and variance as calculated above. The relative frequency of nucleotides will also follow the statistics calculated above. Append the 1000 strings to the end of *yourname.out*.

-The program will then ask the user if they want to process another list.

-If not, the program will exit. If so, the program will prompt for the name of the next file, process it, and append the results to the output file.

Hints

To generate the length of a string from a Gaussian distribution with mean μ and variance σ^2, you can use the `rand()` function (normalized – see RAND_MAX) to generate 2 random numbers, *a* and *b*, uniformly distributed in [0,1). Using the Box-Muller transform it is then possible to compute a random variable *C*, such that:

$$C = \text{sqrt}(-2 \ln(a)) * \cos(2\pi b)$$

Here *C* is a standard Gaussian with mean 0 and variance 1. You can then convert to a normal random variable *D* with mean μ and variance σ^2 as follows:

$$D = \sigma C + \mu$$

Note that here we use the standard deviation, σ, which as you know from basic statistics is simply the square root of variance.

The Rules

- You may **NOT** use any non-primitive data structures to do the math. (No arrays, Vectors, Lists, etc) Just use individual primitive variables (int, double, etc) and std strings. Hopefully this will convince you that data structures make programs more

efficient and easier to write. (Though I suspect you know this already…) Of course, to do the file processing you may use any of the C++ IO classes.

- For this assignment, you must work individually. You may work in pairs in future assignments, but I want you all to write some code on your own first.
- Develop using any IDE you want, but make sure your code runs correctly with g++. Make sure to provide a Makefile so I can build your code easily. (See "Assignments" for a sample Makefile.)
- Feel free to use whatever textbooks or Internet sites you want to refresh your memory with C++ IO operations, just cite them in a README file turned in with your code. All code you write, of course, must be your own.

Due Date

This assignment is due at 11:59pm on 9-14-2019. Submit it your solution to github following the submission instructions discussed in class. We'll spend a few minutes discussing your experience with this assignment in class.

Grading

Grades will be based on correctness, adherence to the guidelines, and code quality (including the presence of meaningful comments). An elegant, OO solution will receive much more credit than procedural spaghetti code. I assume you are familiar with the standard style guide for C++, which you should follow. (See the course page on Blackboard for a C++ style guide.)