

CS250 Assignment 1

DNA Similarity

Date assigned: Friday, February 5, 2016

Date due: Monday, February 15, 2016

Points: 25

Goals for Assignment 1

1. Reacquaint yourself with Visual Studio and Grace.
2. Implement a C++ program that uses files, 1D & 2D arrays, and character processing.
3. Break up a program into well-defined functions. It is important that your functions be small, and focused on a particular task.

The purpose of this assignment is to determine which pair of DNA strands are the most similar using the P Stat.

The P Stat between two DNA strings (s_1, s_2) is the proportion of corresponding bases that differ between the two strings.

For example, if

- S1 is: AAAACCCGGT
- S2 is: AAACCCGGGT

Then the P Stat, denoted $p(s_1, s_2)$ is $2/10$ or 0.2 . The differing bases are the 4th and 7th ones.

Write a complete C++ program that will read in a collection of at most 20 equal length (at most 80 bases) DNA strands from a file called `dnastrings.txt`. You will read in one DNA strand from the keyboard. This keyboard DNA strand will contain the same number of bases as each of the DNA strands in the file.

Compare each DNA strand in the file to the keyboard DNA strand and display the pair's P Stat.

Finally, display the DNA strand from the file that is most similar to the keyboard DNA strand. If multiple DNA strands produce the same minimum P Stat, choose the DNA strand that appears first in the file.

Input: The file named (`dnastrings.txt`) contains an unknown number of DNA strings. There will be at most 20 strings and the length of each string will not exceed 80 bases. Each string will be separated by a new line. The last DNA strand in the file must be followed by exactly one newline.

```
AAAAAAAAA
CCCCCCCC
TTTTTTTT
```

Your **output** is to look exactly like the following:

```
DNA (length=8) :  
AAAACCCC
```

```
S0 : AAAAAAAAA  
S1 : CCCCCCCC  
S2 : TTTTTTTT
```

```
Keyboard: AAAACCCC  
File     : AAAAAAAAA  
P Stat   : 0.50
```

```
Keyboard: AAAACCCC  
File     : CCCCCCCC  
P Stat   : 0.50
```

```
Keyboard: AAAACCCC  
File     : TTTTTTTT  
P Stat   : 1.00
```

Most Similar:

```
Keyboard: AAAACCCC  
File     : AAAAAAAAA  
P Stat   : 0.5
```

Notes:

1. Your main function is to be mainly variable declarations and function calls.
2. Test your program one function at a time.

To complete this assignment you must submit the following:

1. An electronic copy of your program on Grace

- a) Create a new C++ solution in Visual Studio 2013. Your solution must be called **PUNetID-Assignments**. For example, mine would be called will4614-Assignments. This solution will hold all of your assignment projects for CS 250. You are starting over for CS250, meaning do not include any projects from CS150.
- b) Your project for this assignment must be named **01_DNA**. It is vital that you name your solution and your project correctly!
- c) Type your program (fully documented/commented) into the project. You need to follow the coding standards from the CS250 Web page. These coding standards have been modified to include additional C++ language features introduced in CS250, so please be sure to read the new coding standards.
- d) Pay attention to the example output. Your program's output must look **exactly** like the sample output. The spacing and newlines in your output must match exactly.

- e) Make sure that your program builds without errors & warnings and runs correctly. If you get any errors or warnings, double check that you typed everything correctly. Be aware that C++ is case-sensitive. You will lose 10% if there are any warnings and 70% if your program does not build successfully.
- f) Once you are sure that the program works, it is time to submit your program. You do this by logging on to Grace and placing your complete solution folder in the **CS250-0X Drop** folder.
- g) The solution must be in the drop folder by the time class starts on the day the assignment is due. Anything submitted after that will be considered late.

2. A hard copy of your program

- a) The hard copy must be placed on the instructor's desk by the time class starts on the day that it is due.
- b) The hard copy must be printed in color, double-sided, and stapled in the upper left corner if your solution contains multiple pages.
- c) Your tab size must be set to 2 and you must not go past column 80 in your output.

**Remember, if you have any problems, come to me straight
away with your project on Grace. Good Luck!!!! 😊**

Hint: You might review lab 13 and assignment 7 from CS 150.