CS123A Module 2 Week 6
Programming Assignment #1 (40pts) and
Programming Assignment #2 (40pts)

PROGRAMMING ASSIGNMENT #1:

NON-CS MAJORS:

Instructions:

1. Use your favorite genome portal to locally align (using BLAST) the sequences in the files seq1.txt and seq2.txt located in Files -> Module 2 Alignment -> Week 6 -> Data folder. Report the score that you obtain
2. Although we have not discussed using the NCBI portal to perform BLAST-global alignment. You can do this at https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch&PROG_DEF=blastn&BLAST_PROG_DEF=blastn&BLAST_SPEC=GlobalAln&LINK_LOC=BlastHomeLink

   Globally align the sequences in the files seq1.txt and seq2.txt Report the score.

3. Which alignment, global or local, performed the best?
4. Use tools at the NCBI portal (you should now know or have an idea how to do this without being told which specific tool to use) to identify the type (i.e., name) of protein and organisms represented by the sequences in seq1.txt and seq2.txt

CS MAJORS:

This programming assignment involves introducing the Biopython package and using it to perform pairwise sequence alignment to determine if two sequences from two different organisms are similar to each other.

About Biopython:

Biopython is a set of libraries to provide the ability to deal with ”things” of interest to biologists working on the computer. In general this means that you will need to have at least some programming experience (in python, of course!) or at least an interest in learning to program. Biopython's job is to make your job easier as a programmer by supplying reusable libraries so that you can focus on answering your specific question of interest.

One thing to note about Biopython is that it often provides multiple ways of doing the same thing. This can be frustrating since one often wants or needs to know just one correct and efficient way to do something. However, having multiple ways to accomplish the same task can be a real benefit because it gives you lots of flexibility and control over the libraries.

Instructions:

1. If Biopython is not already installed in Python 3.x on your computer, then in a shell window, execute the following command to install it.

   ```
   pip3 install biopython
   ```

2. Read instructions about performing pairwise local and global alignment using Biopython at
   http://biopython.org/DIST/docs/api/Bio.pairwise2-module.html

3. Create a Python module named <your initials>_pairwise_alignment.py that reads in the protein sequences in the files named seq1.txt and seq2.txt and both globally and locally aligns them using the BLOSUM62 matrix. Make sure to use the "`format_alignment`" function in the Biopython package to print out the alignment results.
4. Which alignment, global or local, performed the best?
5. Use tools at the NCBI portal (you should now know or have an idea how to do this without being told which specific tool to use) to identify the type

(i.e., name) of protein and organisms represented by the sequences in seq1.txt and seq2.txt

PROGRAMMING ASSIGNMENT #2:

NON-CS MAJORS:

Do the same as in Programming Assignment #1, except click on the "Algorithm parameters" link and find a Gap Cost that improves, if at all, the score you achieved in Programming Assignment #1.

CS MAJORS:

Do the same as in Programming Assignment #1, except experiment with setting the `penalize_extend_when_opening` and `penalize_end_gaps` alignment parameters to see if you can get a better global alignment score than you achieved in Programming Assignment #1.