CSCI 1300 Introduction to Programming [Fall 2014]
Instructor: Boese
Assignment #7
Due: Fri **Oct 24, by 10am**.

DNA Sequences

Objectives

- Read in command-line arguments
- Read in data from files
- Loop through strings

Measuring DNA Similarity

DNA is the hereditary material in human and other species. Almost every cell in a person's body has the same DNA. All information in a DNA is stored as a code in four chemical bases: adenine (A), guanine (G), cytosine (C) and thymine(T). Different order of these bases means different information.

One of the challenges in computational biology is determining what the *codons* in a DNA sequence represent. A *codon* is a sequence of three DNA or RNA nucleotides that corresponds with a specific amino acid or stop signal during protein synthesis. For example, given the sequence GGGA, the codon could be a GGG or a GGA, depending on where the gene begins and the active bases in the gene. Clues about how to interpret a DNA sequence can be found by comparing an unknown DNA sequence to a known sequence and measuring their similarity. If sequences are similar, then it can be hypothesized that they have similar functions and proteins.

Hamming distance and similarity between two strings

Hamming distance is one of the most common ways to measure the similarity between two strings of the same length. Hamming distance is a position-by-position comparison that counts the number of positions in which the corresponding characters in the string are different. Two strings with a small Hamming distance are more similar than two strings with a larger Hamming distance.

Example:
first string = "AGCCT"
second string = "AACCG"

```
A G C C T
| * | | *
A A C C G
```

In this example, there are three matching characters and two mismatches, so the Hamming distance is two.

The similarity score is then calculated as follows:

similarity_score = (string length - hamming distance) / string length

So for the example provided,

similarity_score = (5-2)/5 → 0.6

Example 1
mouseDNA1.txt

```
CATCATCATCATCATCATCTTTTT
```

humanDNA1.txt

```
ATGCATCATCATCATCATCTTTTT
```

unknownDNA1.txt

```
CATCATCTTCATCATCATCTTTTT
```

Output:

```
MouseCompare = 0.958333
HumanCompare = 0.833333
mouse
```

Example 2
(Each file should have the data all on one line in the file)

mouseDNA2.txt

```
CGCAATTTTTACTTAATTCTTTTTCTTTTAATTCATATATTTTTAATATGTTTACTATTAATGGTTATCATTCACCA
TTTAACTATTTGTTATTTTGACGTCATTTTTTTCTATTTCCTCTTTTTTCAATTCATGTTTATTTTCTGTATTTTTG
TTAAGTTTTCACAAGTCTAATATAATTGTCCTTTGAGAGGTTATTTGGTCTATATTTTTTTTTCTTCATCTGTATTT
TTATGATTTCATTTAATTGATTTTCATTGACAGGGTTCTGCTGTGTTCTGGATTGTATTTTTCTTGTGGAGAGGAAC
TATTTCTTGAGTGGGATGTACCTTTGTTCTTG
```

humanDNA2.txt

```
CGCAAATTTGCCGGATTTCCTTTGCTGTTCCTGCATGTAGTTTAAACGAGATTGCCAGCACCGGGTATCATTCACCA
TTTTTCTTTTCGTTAACTTGCCGTCAGCCTTTTCTTTGACCTCTTCTTTCTGTTCATGTGTATTTGCTGTCTCTTAG
CCCAGACTTCCCGTGTCCTTTCCACCGGGCCTTTGAGAGGTCACAGGGTCTTGATGCTGTGGTCTTCATCTGCAGGT
GTCTGACTTCCAGCAACTGCTGGCCTGTGCCAGGGTGCAGCTGAGCACTGGAGTGGAGTTTTCCTGTGGAGAGGAGC
CATGCCTAGAGTGGGATGGGCCATTGTTCATG
```

unknownDNA2.txt

```
CGCATTTTTGCCGGTTTTCCTTTGCTGTTTATTCATTTATTTTAAACGATATTTATATCATCGGGTTTCATTCACTA
TTTTTCTTTTCGATAAATTTTTGTCAGCATTTTCTTTTACCTCTTCTTTCTGTTTATGTTAATTTTCTGTTTCTTAA
CCCAGTCTTCTCGATTCTTATCTACCGGACCTATTATAGGTCACAGGGTCTTGATGCTTTGGTTTTCATCTGCAAGA
GTCTGACTTCCTGCTAATGCTGTTCTGTGTCAGGGTGCATCTGAGCACTGATGTGGAGTTTTCTTGTGGATATGAGC
CATTCATAGTGTGGGATGTGCCATAGTTCATG
```

Output:

```
MouseCompare = 0.655882
HumanCompare = 0.820588
human
```

Assignment Details:
In this assignment, you will calculate the Hamming distance and similarity scores for sample DNA sequences from a human, a mouse, and an unknown species to determine the identity of the unknown species.

Requirements

- The name of the file must be called **DNA.cpp**
- Comments at the top of your program
  - Your name
  - Date
  - Homework #7
  - Brief description of the assignment (one or two lines max)
- The program takes three command-line arguments, in the following order:

mouseDNA_filename humanDNA_filename unknownDNA_filename

- Use this algorithm to determine the identity of the unknown DNA (if the difference < 0.0001):

```
IF unknownDNA is equally similar to both mouse and human,
        PRINT "identity cannot be determined"
ELSE IF the unknownDNA is more similar to the humanDNA,
        PRINT "human"
ELSE IF the unknownDNA is more similar to the mouseDNA,
        PRINT "mouse"
```

Note: You need to check if they are essentially equal first otherwise this algorithm doesn't work. Do you understand why?

- All data files for a given test case will be the same length.
- You may <u>not</u> create functions, classes, arrays, linked lists, (*e.g., do not use things that we have not covered yet unless you get permission from the instructor first*).
- The output must match exactly to the examples provided (given appropriate inputs).
- If you do not use command-line arguments then you will receive a zero for this assignment.
- Program must be written in C++ and submitted in Moodle.
- Zip the DNA.cpp and submit to Moodle as ***Firstname_Lastname_HW7.zip***.