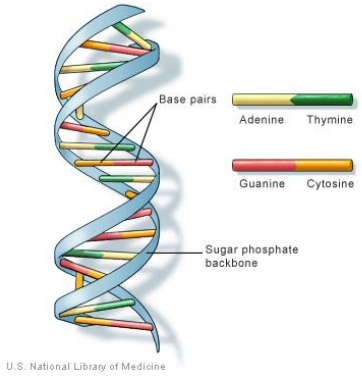
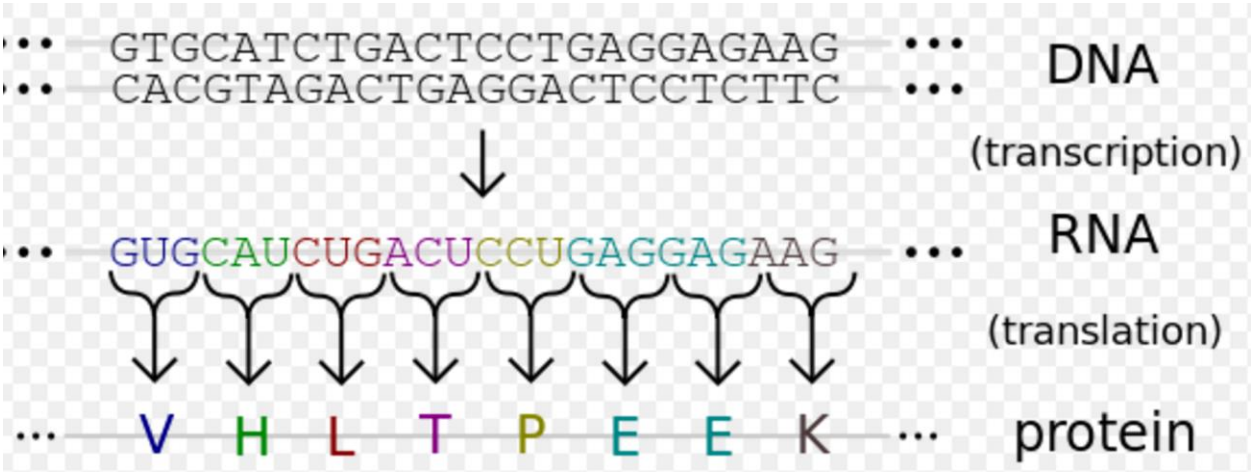
CS250 Assignment 3
Protein Creation

Date Assigned: Wednesday, February 15, 2017
Date Due: Monday, February 27, 2017
Points: 35



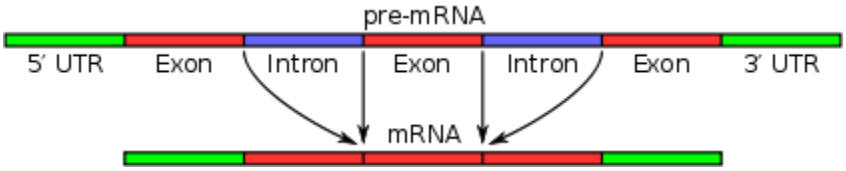
DNA (deoxyribonucleic acid) is hereditary material found in nearly every cell in a person's body and most other organisms. Most DNA is found in the nucleus of the cell. The DNA information is coded from the four chemical bases adenine (A), guanine (G), cytosine (C), and thymine (T). DNA can replicate where each strand of DNA can serve as a pattern for replication.

Genes contain the information that determine a person's traits which are features or characteristics inherited by the person's parents. Each cell in the human body contains around 19,000 genes and the number continues to be revised down over time. In molecular terms, a gene is a region of DNA that contains code used to make a polypeptide (a molecular chain). Polypeptides link together to form proteins which perform a vast array of functions within organisms.



During transcription, a pre-mRNA contains segments called exons and introns. For the purposes of protein translation, the introns are cut out and the exons are sequentially put together. This process, cutting and pasting, is known as splicing. The exons deriving from a gene are collectively referred to as the "coding region." Now how cool is that???

<u>The Problem</u>



Assume the introns and exons of an RNA string have been identified. The problem of creating the coding region is simply deleting the introns and concatenating the exons to form a new string ready for translation. Therefore, given a DNA string (at most 80 characters including the null) and a collection of substrings acting as introns (at most 80 characters including the null), create the protein string. There are an unknown number of introns.

Example

Consider the following sample dataset:

ATGGTGCATCTGACTCCTGAGGAGAAGTAG
CTGACT

The protein string produced is:

MVHPEEK

Question: How does ATG in the original DNA string become M in the protein string?

Answer: Groups of 3 in the coding region are known as codons. Remember, all T's will become U's in the pre-mRNA string; therefore, ATG becomes AUG which is also known as the start codon of a pre-mRNA string and UAA, UAG, and UGA are all stop codons.

How to solve this problem?

1) The file CodonToAminoAcidTable.txt exists in the Public folder on grace. You are to load this table into an array of structs called **asCodonToAminoAcidTable**.

2) The file **strings.txt** contains a DNA string and a collection of substrings acting as introns.

3) Convert the DNA string into a pre-mRNA string.

4) Convert the pre-mRNA string into an mRNA string.

5) Convert the mRNA string into a protein.

The challenge of this problem is the design. Think well defined functions that have one purpose!!!!!

The output from your program is:

```
Protein Creation
----------------

Coding Region
-------------
ATGGTGCATCTGACTCCTGAGGAGAAGTAG

pre-mRNA
AUGGUGCAUCUGACUCCUGAGGAGAAGUAG

mRNA
AUGGUGCAUCCUGAGGAGAAGUAG

Protein
-------
MVHPEEK
```

Notes

1) You must break your program up into .h and .cpp files.

2) Do not use the string data type at all in this assignment. That is, use character arrays only.

To complete this assignment you must submit the following:

1. **An electronic copy of your program on Grace**

a) Add a project called **03ProteinCreation** to your existing solution **PUNetID-Assignments**. It is vital that you name your project correctly!
b) 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.
c) 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.
d) 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 40% if your program does not build successfully.

e) 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 correct drop folder based on the section of the course in which you are enrolled (**CS250-XX Drop**).

f) 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.

g) If you drop multiple solutions, you will lose 10% of the assignment points, so do not drop until you are entirely sure you are completely done working on the assignment.

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 a flash drive or on Grace. Good Luck!!!! □