CSCI 135 – Software Design and Analysis I
Assignment 2

Introduction

This assignment gives you practice programming with classes and file I/O. As with all programs, you are free to discuss with or take inspiration from the instructor, other students and other sources, but your implementation must be your own work.

Your program must be written in C++ and must compile and run on eniac.geo.hunter.cuny.edu. Your program should also follow the guidelines set out here: http://www.compsci.hunter.cuny.edu/~sweiss/course_materials/csci235/programming_rules.pdf. Submit the source code for your program as one (zipped) file via Blackboard.

Submission and Grading

Submit your source code (.cpp and .hpp files) for your implementation on Blackboard. Your program will be graded as follows:
50% = Correctness
15% = Performance
15% = Design
10% = Style
10% = Documentation

Assignment

Background

A DNA string, also called a DNA strand, is a finite sequence consisting of the four letters a, c, g, and t in any order. The four letters stand for the four nucleotides: adenine, cytosine, guanine, and thymine. Nucleotides, which are the molecular units from which DNA and RNA are composed, are also called bases.

A special enzyme called RNA polymerase uses the information in DNA to create RNA. The process of creating RNA from DNA is called transcription. A RNA string or RNA strand is a finite sequence consisting of the four lowercase letters a, c, g, and u. The a, c, and g have the same names as they do in DNA, but the u represents uracil. When DNA is transcribed to RNA by RNA polymerase, each thymine base is converted to uracil. Hence RNA strings have “u”s wherever DNA has “t”s.

RNA in turn serves as a template for the construction of proteins, which are sequences of amino acids. Proteins are synthesized within the ribosomes of living cells by a process called translation. In translation, the RNA string is viewed as a sequence of three-letter groups called codons. Each codon codes for a particular amino acid. For example, guu codes for valine, and uca codes for cysteine. There are 64 different codons. On the other hand, there are only 20 different amino acids. Some amino acids

are coded for by multiple codons. For example, uca, ucc, ucg, and ucu all code for cysteine. Some codons do not code for any amino acids; they are stop codes. There are three stop codons: uaa, uag, and uga. Stop codes are used during protein synthesis to terminate reading of the RNA string. Not all of an RNA string is translated into protein; there are large regions that act like gaps. As the RNA is read, when a gap is reached, it is skipped over until a special start codon is found that tells the ribosome to begin creating amino acids again. When it sees a stop codon it stops and keeps reading until it finds another start codon, and so on, until the entire strand is read. As an example, the RNA strand augguuuauggucucuga is read as the following sequence of codons aug guu uau ggu cuc uga. Consulting a table of these mappings, we see that aug is a start codon that codes for methionine (Met), guu, for valine (Val), uau, for tyrosine (Tyr), ggu, for glycine (Gly), cuc, for leucine (Leu), and uga is a stop codon. Therefore, the sequence Met-Val-Tyr-Gly-Leu is created from this RNA fragment.

Below is a table of codons and the three-letter name for the corresponding amino acids.

Codon	Amino Acid	Codon	Amino Acid	Codon	Amino Acid	Codon	Amino Acid
uuu	Phe	ucu	Ser	uau	Tyr	ugu	Cys
uuc	Phe	ucc	Ser	uac	Tyr	ugc	Cys
uua	Leu	uca	Ser	uaa	TER	uga	TER
uug	Leu	ucg	Ser	uag	TER	ugg	Trp
cuu	Leu	ccu	Pro	cau	His	cgu	Arg
cuc	Leu	ccc	Pro	cac	His	cgc	Arg
cua	Leu	cca	Pro	caa	Gln	cga	Arg
cug	Leu	ccg	Pro	cag	Gln	cgg	Arg
auu	Ile	acu	Thr	aau	Asn	agu	Ser
auc	Ile	acc	Thr	aac	Asn	agc	Ser
aua	Ile	aca	Thr	aaa	Lys	aga	Arg
aug	Met	acg	Thr	aag	Lys	agg	Arg
guu	Val	gcu	Ala	gau	Asp	ggu	Gly
guc	Val	gcc	Ala	gac	Asp	ggc	Gly
gua	Val	gca	Ala	gaa	Glu	gga	Gly
gug	Val	gcg	Ala	gag	Glu	ggg	Gly

You may also find the full name and molecular formula for each amino acid from wikipedia (http://en.wikipedia.org/wiki/Amino_acid) or from the attached file, which your program must use.

Assignment Requirements

Write a program to perform the following:

1. Read the file of codons (attached to this assignment)
2. Read a DNA strand from the user which should contain a, c, g and t (without spaces)
3. Transcribe the DNA strand to an RNA strand (see below)

Note on "Transcribe the DNA strand to an RNA strand":

The user will input a DNA strand (which should contain only the letters a,c,g, and t without spaces or newline characters) and the program should display the sequences of three letter amino acids. If a stop codon is reached, the next sequence of amino acids should be displayed on a new line. For example, given a DNA strand that transcribes to the following RNA strand:

augguuuauggucucugaauuaaucuccauguuuuaucacuaa

... the output of the program must be:

Met-Val-Tyr-Gly-Leu
Met-Phe-Tyr-His

The above RNA strand can be broken into three parts:

1. **aug guu uau ggu cuc uga** - aug is a start codon so we output a sequence of amino acids (Met-Val-Tyr-Gly-Leu) until we see a stop codon (uga).
2. **auu aau cucc** - this is a gap. Note that the gap is *not* necessarily a multiple of three.
3. **aug uuu uau cac uaa** - at this point we see the start codon aug so we output another sequence (Met-Phe-Tyr-His) until we see a stop codon (uaa).

After the transcription, the program must also display the following information for each amino acid that that was part of a sequence:

- Three letter name
- Full Name
- Molar mass

Design & Implementation

Your implementation may use a class called AminoAcid, representing the following properties of an amino acid:

- Name
- Three-letter name
- Molecular Formula
- Codons

The class may have accessor and mutator functions to get or set amino acid properties.