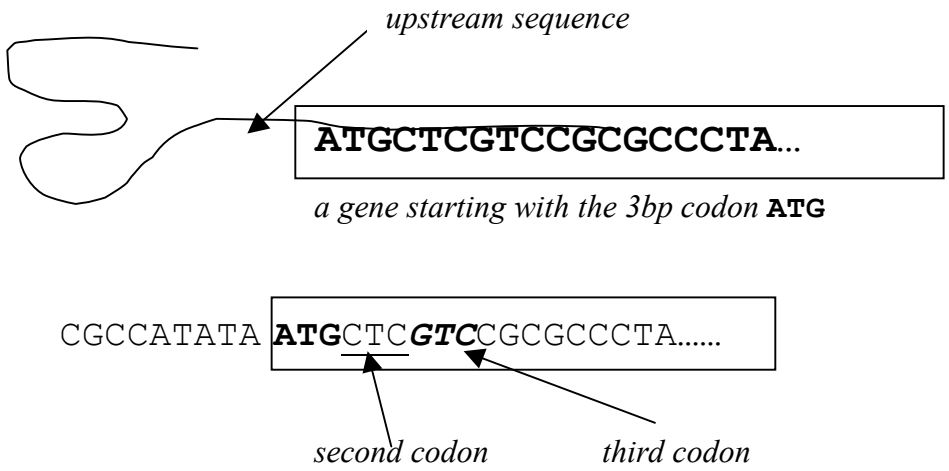
DNA

Playing with Strings

Write a Perl program to print a report about a specific DNA sequence:

cgccatataatgctcgtccgcgcccta

The sequence (starting in lowercase) is a "snip" of DNA that includes some upstream (intergenic) sequence and the beginning of a gene (genic sequence). Since all genes begin with the three nucleotide code ATG, we can "easily" find the start of a gene. All the sequence to the left (upstream) of ATG is the upstream (intergenic – between the gene) sequence and the sequence to the right (and including) ATG is the genic (gene) sequence. Note: the entire gene sequence is not included in this small sequence.



Within a gene, every three nucleotides is called a codon. ATG is the initial codon in every gene, but other codons vary depending on the particular gene. In this example, CTC is the second codon and GTC is the third codon and so on.

Your task is to locate and report the position of the ATG codon within the starting sequence and then take substrings (`substr`) of the entire sequence to form two other strings that hold the upstream and genic sequences. In addition to printing out the starting sequence and length, locations and three-character codes of the first three codons in the gene, upstream and genic sequences and their respective lengths (see a sample report below), your program will also print out the reverse complement of the genic sequence. Note: your program should *not* only work on this sequence but on any sequence that contains ATG, therefore, use variables (e.g., `$positionATG`) rather than 9.

Finally (once you have all of the above working), print the original sequence with only the ATG start codon in capital letters (uppercase) while the rest of the sequence is in lowercase. Also, to try and get a feel for whether a region is AT-rich, print a count of the number of As and Ts in the upstream region.

A **sample output** is shown below. Your program's output need not be exactly identical but you should print out the information in the same order and obviously your answers should agree.

```
++++++++++++ Upstream and Genic Report ++++++++++++++++++++++++

Starting sequence is:    cgccatataatgctcgtccgcgcccta
Converted to uppercase:  CGCCATATAATGCTCGTCCGCGCCCTA

Length of starting sequence is: 27
-----------------------------------------------------------------

ATG start codon begins in position (bp) 10
    followed by codon CTC in position (bp) 13
    followed by codon GTC in position (bp) 16
-----------------------------------------------------------------

Upstream sequence is:  CGCCATATA

Upstream length (bp): 9

-----------------------------------------------------------------

Gene sequence is:  ATGCTCGTCCGCGCCCTA

Gene length (bp): 18

-----------------------------------------------------------------

Gene + Strand:  ATGCTCGTCCGCGCCCTA

Gene - Strand:  TAGGGCGCGGACGAGCAT

: you complete the last steps ….
-----------------------------------------------------------------
```

This program is **due on Monday September 17** (specifically, dropped to Bb by 5am on Tuesday, Sept 18). You must submit your Perl program via Blackboard (Bb). Also slide a landscaped printout of your program under Professor LeBlanc's or Dyer's door by class on Tue Sept 18. No late submissions are accepted. (Read that last sentence again).

Submit:
(1) one hardcopy of your Perl program (use a good name, e.g., **`LeBlanc_a1.pl`**)
(2) one page that shows the **OUTPUT** of your program.
(3) **You** *must* staple your Perl (`LeBlanc_a1.pl`) and your output together.

Here is some help getting started ….

Your program must have some documentation at the top in a "box".

```
#!/usr/bin/perl

use strict;
use warnings;

#=============================================================
#
# Summary: This Perl program isolates the upstream and genic
#          regions of a sequence. A report is printed, a sample
#          of which is shown below:
#
#          (you paste a sample of your program's output here)
#
# Programmer:  Ima Intergenic
#
# Date Last Modified:
#  09/10/2007 -- started program, finished length of sequence
#  09/11/2007 -- trouble with getting correct location of ATG
#  09/12/2007 -- fixed ATG location, finished program
#
#=============================================================

print "+++++++++ Upstream and Genic Report +++++++++++++++\n\n";

my $someSequence;       # upstream and start of a gene ...

$someSequence     = "cgccatataatgctcgtccgcgcccta";

print "Starting sequence is:    $someSequence \n";

# convert all nucleotides to uppercase
:

print "Converted to uppercase:  $someSequence \n\n";

:
print "Length of starting sequence is: $seqLength \n";

print "-----------------------------------------------------\n\n";


# get the position of the start codon "ATG"
my $ATGPosition;

$ATGPosition = index( ……
:

print "-----------------------------------------------------\n\n";

my $upStream;
$upStream = substr( ……

print "Upstream sequence is:  $upStream \n\n";
:
```