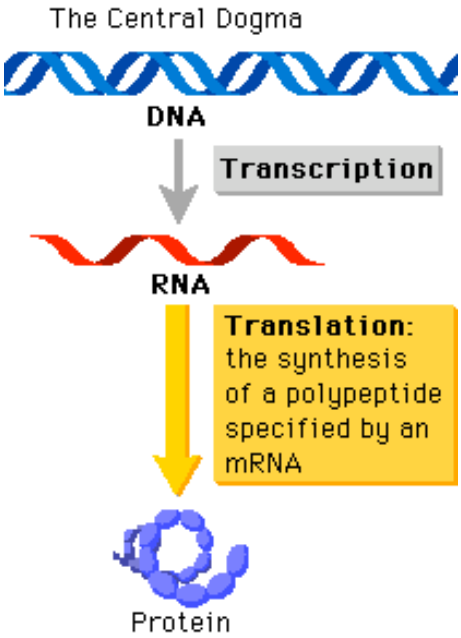
DNA

mini Gene Finder

The Central Dogma of molecular biology relates DNA, RNA, and proteins. Briefly put, the DNA sequence (gene) provides instructions for the protein's production.



STEP ONE: A DNA sequence such as:

`AGTCTGACCTAG`

is **TRANSCRIBED** to a messenger RNA (mRNA):

`UCAGACUGGAUC`.

Remember that A complements T and C complements G in DNA. However, when making a complementary RNA (transcribing), `U` (uracil) is used instead of `T`.

STEP TWO: The mRNA is **TRANSLATED** into PROTEIN (strings of amino acids). Special translation structures (ribosomes) "read" the mRNA three bases at a time. Every "triplet" codes for a particular amino acid except for three triplets that code for "STOP" at the end of a gene.

Important Rules when building your program:

1. The first triplet in an mRNA is "`AUG`" which also codes for the amino acid methionine (Met). The triplet `AUG` can be BOTH a "Start" and a simple instruction for putting a methionine somewhere in the middle of a protein. YOUR PROGRAM should search for Met *both* at the beginning and elsewhere and report the whereabouts.

2. The last triplet should be one of the stops, indicated by three DOTS in this code (`...`). Every mRNA should have a stop at the end but not elsewhere. YOUR PROGRAM should search for STOP both at the end and elsewhere and report on the whereabouts.

3. An mRNA strand may be "read" in any of three "**reading frames**". One of those is often better than the other two. "Better" means it has a proper START and STOP and no STOPS elsewhere. YOUR OUTPUT will help you determine whether ANY of the frames looks like a good gene.

You can fetch a "Starter Kit" (.zip file) for this program. The kit contains the specification ("spec", this file) and a template of a Perl program that'll get you started. The template includes some Perl subroutines (`sub`), sometimes called functions, that we have written or started to help with your solution. Click on the link for Chapter 6 "**If …Then …Else Statements"** found at this URL:

`http://cs.wheatoncollege.edu/mleblanc/dna`

In particular, we've given you (or started) three functions to perform the following tasks:

```
# these subroutines appear at the bottom of your starting code


#-------------------- Subroutines for you to use ----------------------
#
# These subroutines (sometimes called functions) below are ready for
# you to use in your program. Think of these routines as "buttons on
# your calculator", they are ready to use. When you "use" one of these
# routines, we say "your program is CALLING" for the use of that routine.
```

getAminoAcidLookUpTable	builds a Perl hash table to convert three nucleotides (e.g, "AUG") to the corresponding three letter amino acid abbreviation (e.g., "Met").
transcribe	returns the complement of a strand of DNA and substitutes U for T.
translate	converts a string of nucleotides in a messenger RNA (mRNA) into a string of three-letter-codes for each amino acid; every three nucleotides (a codon) is converted into one amino acid abbreviation.

A complete **translation** table that we will use in this program is shown below.

Amino acid symbol	Triplets of nucleotides that code for the amino acid
Ala	GCU GCC GCA GCG
Arg	CGU CGC CGA CGG AGA AGG
Asn	AAU AAC
Asp	GAU GAC
Cys	UGU UGC
Gln	CAA CAG
Glu	GAA GAG
Gly	GGU GGC GGA GGG
His	CAU CAC
Ile	AUU AUC AUA
Leu	UUA UUG CUU CUC CUA CUG
Lys	AAA AAG
Met	AUG
Phe	UUU UUC
Pro	CCU CCC CCA CCG
Ser	UCU UCC UCA UCG AGU AGC
Thr	ACU ACC ACA ACG
Trp	UGG
Tyr	UAU UAC
Val	GUU GUC GUA GUG
...	UAA UAG UGA

A sample output from Betsey's solution:

```
============================================================
             DNA:        TATACATGCGTTTACTATAGCATTTAAGGATT
                        (-- transcribe --)
             RNA:        AUAUGUACGCAAAUGAUAUCGUAAAUUCCUAA

============================================================
RNA length:              32
             RNA:        AUAUGUACGCAAAUGAUAUCGUAAAUUCCUAA
                               (-- translate --)
READING FRAME (1):       IleCysThrGlnMetIleSer...IlePro

There is no Met at the start.
There is no stop at the end.
There is at least one Met somewhere other than start.
There is at least one stop interrupting the gene
============================================================
RNA length:              31
             RNA:        UAUGUACGCAAAUGAUAUCGUAAAUUCCUAA
                               (-- translate --)
READING FRAME (2):       TyrValArgLys...TyrArgLysPheLeu

There is no Met at the start.
There is no stop at the end.
There are no Mets in positions outside of the start site.
There is at least one stop interrupting the gene
============================================================
RNA length:              30
             RNA:        AUGUACGCAAAUGAUAUCGUAAAUUCCUAA
                               (-- translate --)
READING FRAME (3):       MetTyrAlaAsnAspIleValAsnSer...

There is a met at the start site.
There is a stop at the end.
There are no Mets in positions outside of the start site.
There are no interrupting STOPS.
============================================================
```