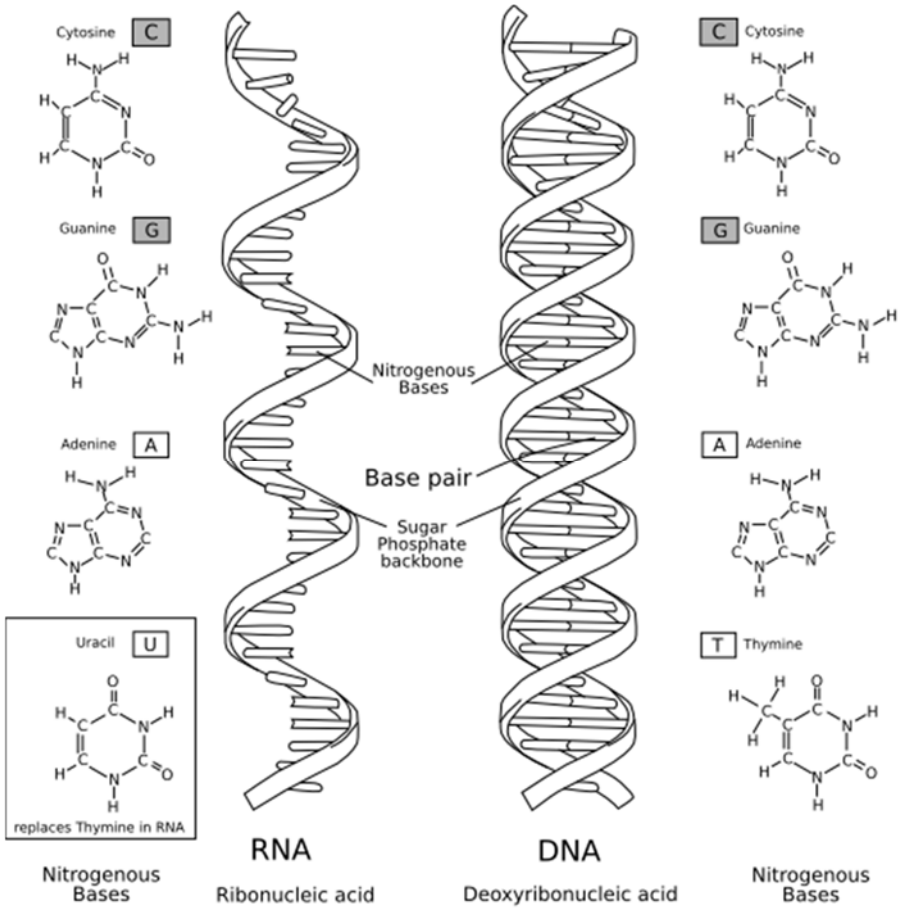
CISC 1115 (Science Section)
Brooklyn College
Professor Langsam

Assignment #6

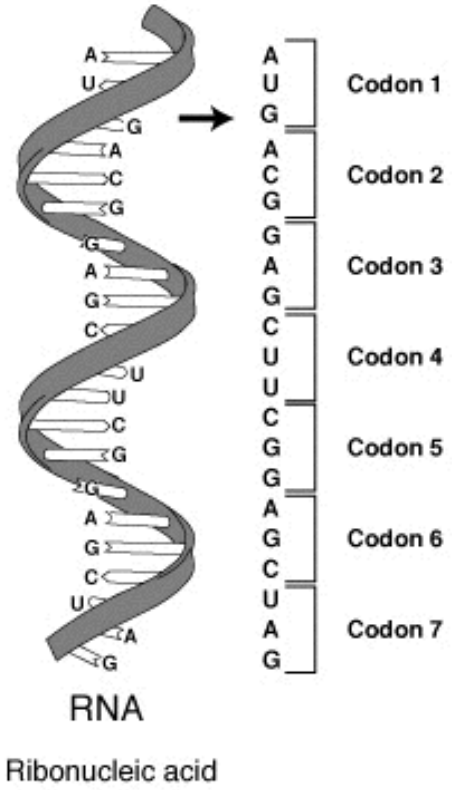
The Genetic Code[1]

Deoxyribonucleic acid, or **DNA**, is a molecule that contains the instructions used in the development and functioning of all known living organisms. The main role of DNA is the long-term storage of information and it is often compared to a set of blueprints, since DNA contains the instructions needed to construct other components of cells, such as proteins and RNA molecules. DNA is a molecule in the form of a double helix (twisted-ladder) Attached to the backbone are of four types of molecules called bases – these form the rungs of the ladder. It is the sequence of these four bases along the backbone that encodes information. The four bases found in DNA are adenine (abbreviated A), cytosine (C), guanine (G) and thymine (T). **RNA** is similar with the base uracil (U) rather than thymine. See the figure below.



[1] http://en.wikipedia.org/wiki/Genetic_code

The genome of an organism is inscribed in the DNA. The portion of the genome that codes for a protein or an RNA molecule is referred to as a gene. The **genetic code** is the set of rules by which information encoded in genetic material (DNA or RNA sequences) is translated into proteins (amino acid sequences) by living cells. Specifically, the code defines a mapping between tri-nucleotide sequences called **codons** and amino acids; every triplet of nucleotides in a nucleic acid sequence specifies a single amino acid



Since there are 4 possible bases (A, C, G, T or U) and each codon consists of 3 bases there are $4^3 = 64$ different combinations possible with a triplet codon of three nucleotides. If, for example, an RNA sequence, UUUAAACCC is considered, there are three codons, namely, UUU, AAA and CCC, each of which specifies one amino acid. This RNA sequence will be translated into an amino acid sequence, three amino acids long.

The standard genetic code is shown in the following tables. Table 1 shows what amino acid each of the 64 codons specifies. Table 2 shows what codons specify each of the 20 standard amino acids involved in translation. These are called forward and reverse codon tables, respectively. For example, the codon AAU represents the amino acid asparagine, and UGU and UGC represent cysteine (standard three-letter designations, Asn and Cys respectively). Note several codons can code for the same amino acid.

Table 1: RNA Codon table - This table shows the 64 codons and the amino acid each codon codes for.

		2nd base			
		U	**C**	**A**	**G**
1st base	**U**	UUU (Phe/F)Phenylalanine UUC (Phe/F)Phenylalanine UUA (Leu/L)Leucine UUG (Leu/L)Leucine	UCU (Ser/S)Serine UCC (Ser/S)Serine UCA (Ser/S)Serine UCG (Ser/S)Serine	UAU (Tyr/Y)Tyrosine UAC (Tyr/Y)Tyrosine UAA Ochre (*Stop*) UAG Amber (*Stop*)	UGU (Cys/C)Cysteine UGC (Cys/C)Cysteine UGA Opal (*Stop*) UGG (Trp/W)Tryptophan
	C	CUU (Leu/L)Leucine CUC (Leu/L)Leucine CUA (Leu/L)Leucine CUG (Leu/L)Leucine	CCU (Pro/P)Proline CCC (Pro/P)Proline CCA (Pro/P)Proline CCG (Pro/P)Proline	CAU (His/H)Histidine CAC (His/H)Histidine CAA (Gln/Q)Glutamine CAG (Gln/Q)Glutamine	CGU (Arg/R)Arginine CGC (Arg/R)Arginine CGA (Arg/R)Arginine CGG (Arg/R)Arginine
	A	AUU (Ile/I)Isoleucine AUC (Ile/I)Isoleucine AUA (Ile/I)Isoleucine AUG (Met/M)Methionine, *Start*	ACU (Thr/T)Threonine ACC (Thr/T)Threonine ACA (Thr/T)Threonine ACG (Thr/T)Threonine	AAU (Asn/N)Asparagine AAC (Asn/N)Asparagine AAA (Lys/K)Lysine AAG (Lys/K)Lysine	AGU (Ser/S)Serine AGC (Ser/S)Serine AGA (Arg/R)Arginine AGG (Arg/R)Arginine
	G	GUU (Val/V)Valine GUC (Val/V)Valine GUA (Val/V)Valine GUG (Val/V)Valine	GCU (Ala/A)Alanine GCC (Ala/A)Alanine GCA (Ala/A)Alanine GCG (Ala/A)Alanine	GAU (Asp/D)Aspartic acid GAC (Asp/D)Aspartic acid GAA (Glu/E)Glutamic acid GAG (Glu/E)Glutamic acid	GGU (Gly/G)Glycine GGC (Gly/G)Glycine GGA (Gly/G)Glycine GGG (Gly/G)Glycine

Table 2: Inverse table

Ala/A	GCU, GCC, GCA, GCG	**Leu/L**	UUA, UUG, CUU, CUC, CUA, CUG
Arg/R	CGU, CGC, CGA, CGG, AGA, AGG	**Lys/K**	AAA, AAG
Asn/N	AAU, AAC	**Met/M**	AUG
Asp/D	GAU, GAC	**Phe/F**	UUU, UUC
Cys/C	UGU, UGC	**Pro/P**	CCU, CCC, CCA, CCG
Gln/Q	CAA, CAG	**Ser/S**	UCU, UCC, UCA, UCG, AGU, AGC
Glu/E	GAA, GAG	**Thr/T**	ACU, ACC, ACA, ACG
Gly/G	GGU, GGC, GGA, GGG	**Trp/W**	UGG
His/H	CAU, CAC	**Tyr/Y**	UAU, UAC
Ile/I	AUU, AUC, AUA	**Val/V**	GUU, GUC, GUA, GUG
START	AUG	**STOP**	UAG, UGA, UAA

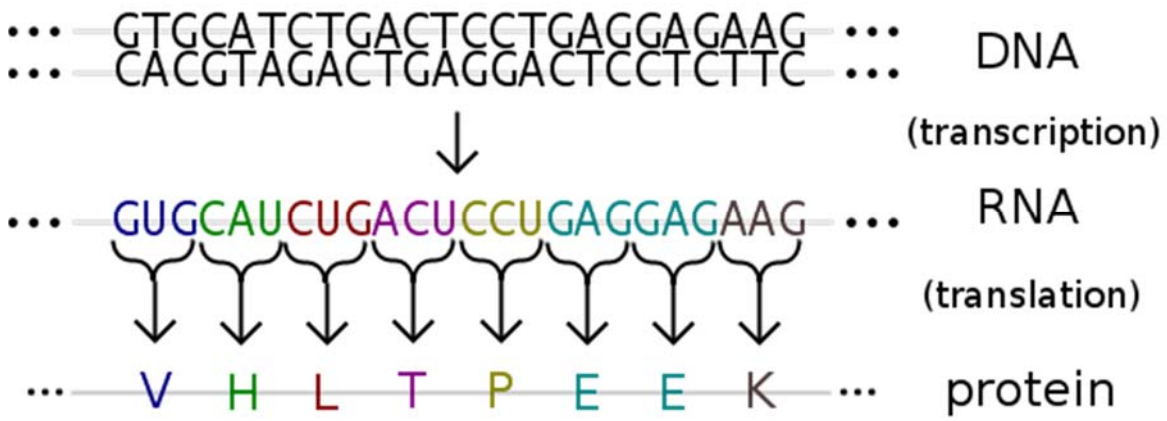
Note that a codon is defined by the initial nucleotide from which translation starts. For example, the string GGGAAACCC, if read from the first position, contains the codons GGG, AAA and CCC; and if

read from the second position, it contains the codons GGA and AAC; if read starting from the third position, GAA and ACC. Partial codons have been ignored in this example. Every sequence can thus be read in three **reading frames**, each of which will produce a different amino acid sequence (in the given example, Gly-Lys-Pro, Gly-Asp, or Glu-Thr, respectively).

The actual frame of a protein is translated in is defined by a **start codon**, usually the first AUG codon in the mRNA sequence. There are three stop codons (UAG, UGA, UAA) which signal then end of a sequence.

Proteins are large organic compounds made of amino acids arranged in a linear chain. The sequence of amino acids in a protein is defined by a gene and encoded in the genetic code. Proteins are essential parts of organisms and participate in every process within cells.

The entire process may be represented by the following diagram:



Write a program that examines a sequence of bases in a strand of RNA and prints the amino acid sequence for each protein coded within that strand. For example, given the following portion of an RNA strand:

…AAUUGU**AUG**AAAUUUCCUGAAUAU**UAG**GAUGCUCAAAAA**AUG**UGGUUUUUGUUGGAACAAGAC**UAA**U ACUUUU…

Your program should print the following:

```
Protein #1: KFPEY
Protein #2: WFLLEQD
```

Note that each protein begins with a *start* codon and ends with a *stop* codon. The sequence in between the stop codon and the next start codon does not code for any protein and is known as **junk DNA**. An actual protein may be hundreds or thousands of amino acids long. Also note that we are using the one letter abbreviations of the 20 amino acids that make up all proteins.

Strategy

1. Create a data file that contains the information given in Table 2. The table should have one entry per line:

GCU	A
GCC	A
GCA	A
GCG	A
CGU	R
CGC	R
...	

2. Write a method, `readRNACodonTable`, that reads this data file into two 1-dimensional arrays. (If you are particulary ambitious you may use a single two dimensional array.)
3. Write a method, `sort`, that sorts this array in codon order.
4. Write a method, `codonLookup`, that given a three character codon returns the one-letter abbreviation for the corresponding amino acid by searching the table you created in step 2.
5. Write a method `isValidRNA` that will receive a character string as a parameter and will return **`true`** if the string represents a RNA strand, and will return **`false`** otherwise. If the character string contains any characters other than A,C,G and U, (even spaces), then it is not a valid RNA strand.
6. The main method should:
    a. Read the RNA strand from a file as a string.
    b. Validate this string.
    c. Scan this string until a *start* codon is encountered.
    d. Continue to scan the RNA strand three characters at a time and call the method `codonLookup` to identify the corresponding amino acid.
    e. Print the one-letter abbreviation returned by `codonLookup.`
    f. Repeat step 5b-d until a *stop* codon is returned.
    g. Continue scanning the RNA strand (steps 5a-d) until the end of the strand is reached.

Extra Credit:

7. Have the main method print the junk DNA sequence.
8. Write a reverse method that accepts an amino acid sequence and prints each of the possible codon sequences that would code this protein. (There may be more than one. *Why*?) – Use this method on each protein that you have found in the original assignment.

Data: (copy this carefully – you may also download it from the web at:
http://eilat.sci.brooklyn.cuny.edu/cisc1115/CISClassPage.htm)

...AACAAUAUUAUGCAACAGUGUCCUCCCUUAUGAGCGUGUGGUUAGAAUUGUAUGAAAUUUCCUGAAU
AUUAGGAUGCUCAAAAAAUGUGGUUUUUGUUGGAACAAGACUAAUACUUUUUUGUUGAUGAGAAUGAA
ACCCCCCAAAUUUAGAGCUGCCAGACAUCAACCCUUUAAACCCCCCUAGUUUCCCAAA...