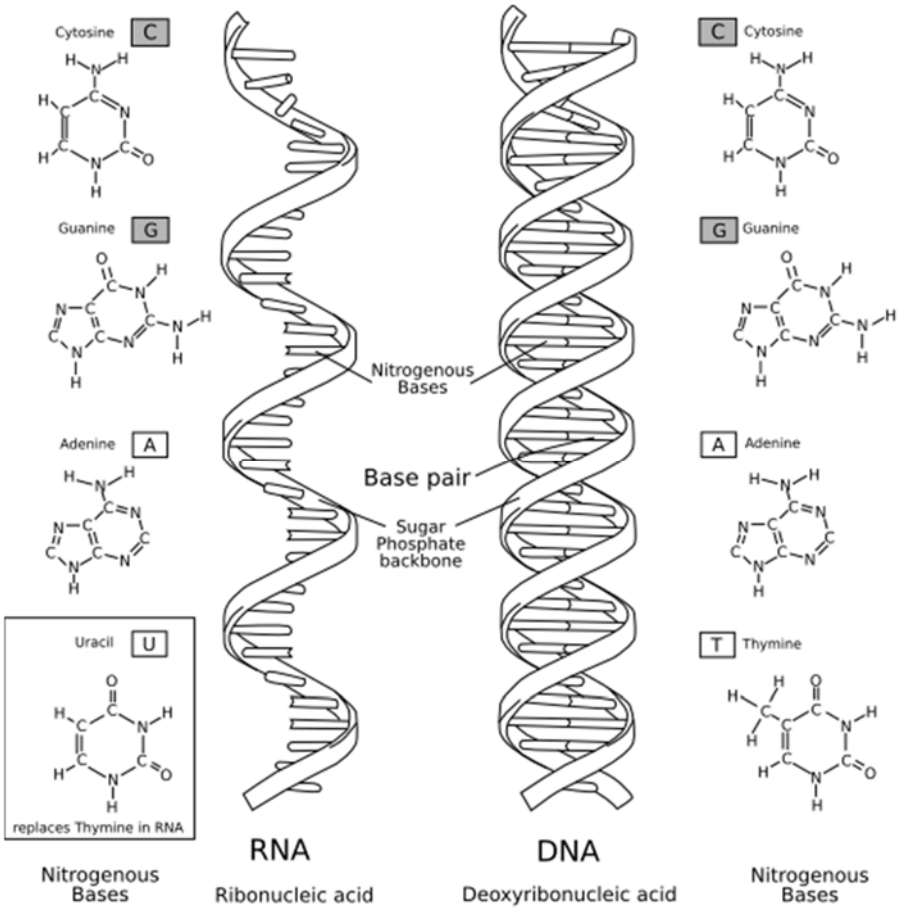
CISC 1115 (Science Section)
Brooklyn College
Professor Langsam

Assignment #6

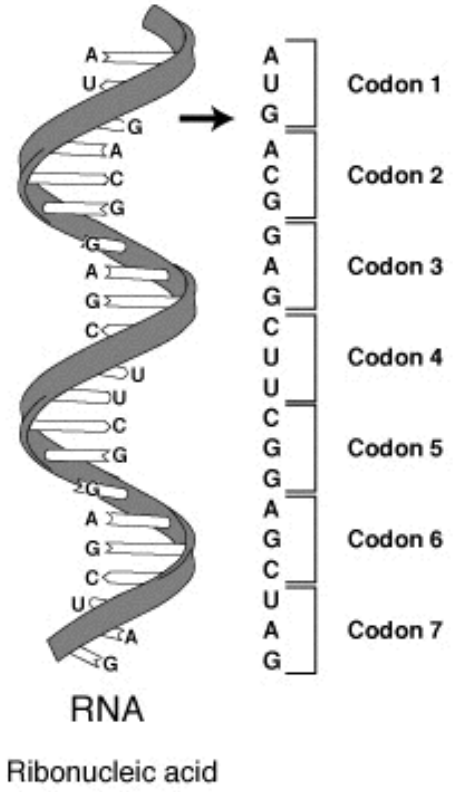
The Genetic Code[1]

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.

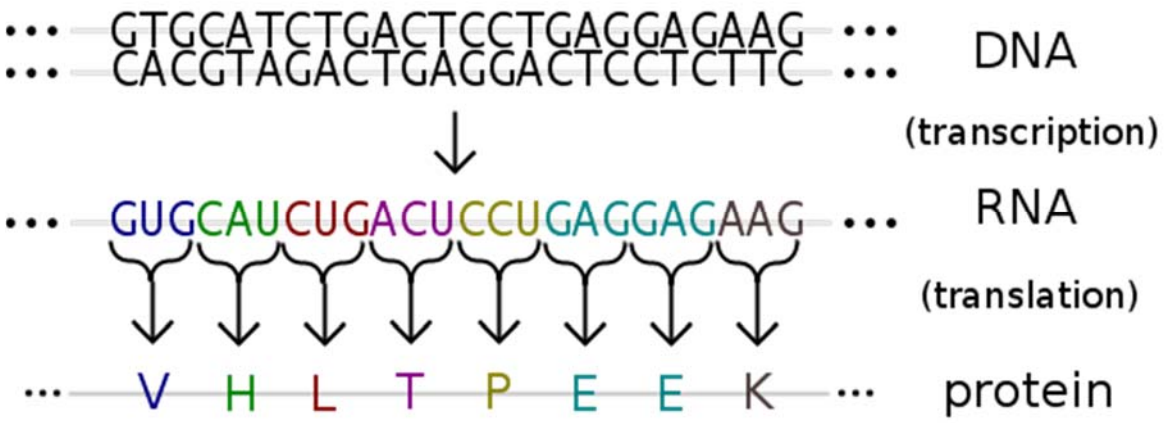
		2nd base			
		U	**C**	**A**	**G**
1st base	**U**	UUU (Phe/F)Phenylalanine UUC (Phe/F)Phenylalanine UUA (Leu/L)Leucine UUG (Leu/L)Leucine	UCU (Ser/S)Serine UCC (Ser/S)Serine UCA (Ser/S)Serine UCG (Ser/S)Serine	UAU (Tyr/Y)Tyrosine UAC (Tyr/Y)Tyrosine UAA Ochre (*Stop*) UAG Amber (*Stop*)	UGU (Cys/C)Cysteine 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

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:



What Makes Us Human?

It is interesting to compare the genome of human beings with that of other species. In a fascinating article that appeared in *Scientific American*, the author writes:

> As a biostatistician with a long standing interest in human origins, I was eager to line up the human DNA sequence next to that of our closest living relative [the common chimpanzee (*Pan Troglodytes*)] and take stock. A humbling truth emerged: our DNA blueprints are nearly 99 percent identical to theirs. That is, of the three billion letters that makes up the human genome, only 15 million of them-less than 1 percent-are [different] [2].

Write a computer program that locates the differences between the human, chimpanzee and chicken portion of genome known as *human accelerated region 1* (HAR1). This part of the genome is believed to be active in a type of neuron that plays a key role in the pattern and layout of the developing cerebral cortex, the wrinkled outermost brain layer[3]. (For the significance of this gene, refer to the referenced *Scientific American* article.)

In addition, you will identify the protein that is coded for by these codons.

[2] Katherine S. Pollard, *Scientific American*, 300(5), May 2009, page 44
[3] Katherine S. Pollard, *Scientific American*, 300(5), May 2009, page 46

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 one-dimensional arrays.
3. Write a method, `sort` that sorts these 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, `difference`, that given two strings (representing gene sequences) prints out the position and identity of each base (ACGT) that differ in the two strings.
6. The main method should:
    a. Read the three DNA strands (human, chimpanzee, chicken) from a file into three strings.
    b. Call a method `isValidDNA` to determine that the strand only contains A C G or T. If it contains any other character, abort the program (*System.exit*(0)).
    c. Call method `difference` three times in order to determine the differences between human-chicken, human-chimpanzee and chimpanzee-chicken HAR1
    d. Reprocess the three DNA strands (human, chimpanzee and chicken HAR1) three bases at a time, calling `codonLookup` in order to determine the corresponding amino acid. Remember that in the RNA codon lookup table, a U is substituted for each T in the DNA strand.

Data: (Download *HumanHNR1.txt* , *ChimpanzeeHNR1.txt* and *ChickenHNR1* from the web at: http://eilat.sci.brooklyn.cuny.edu/cisc1115/CISClassPage.htm)