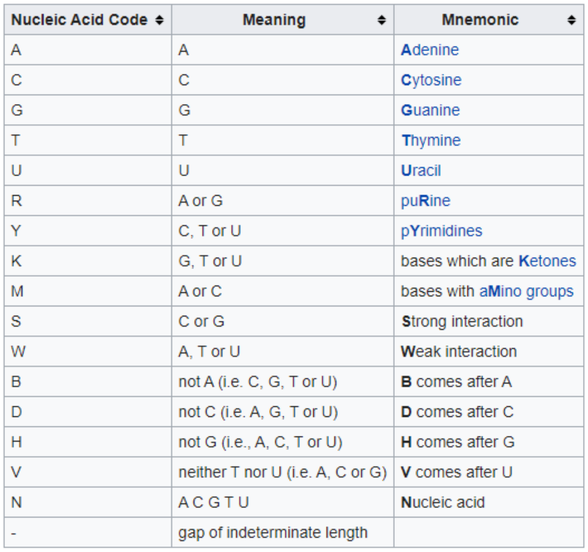
Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics and evolution. They have a set of databases for genetic and protein sequences for various vertebrates here: http://useast.ensembl.org/info/data/ftp/index.html.

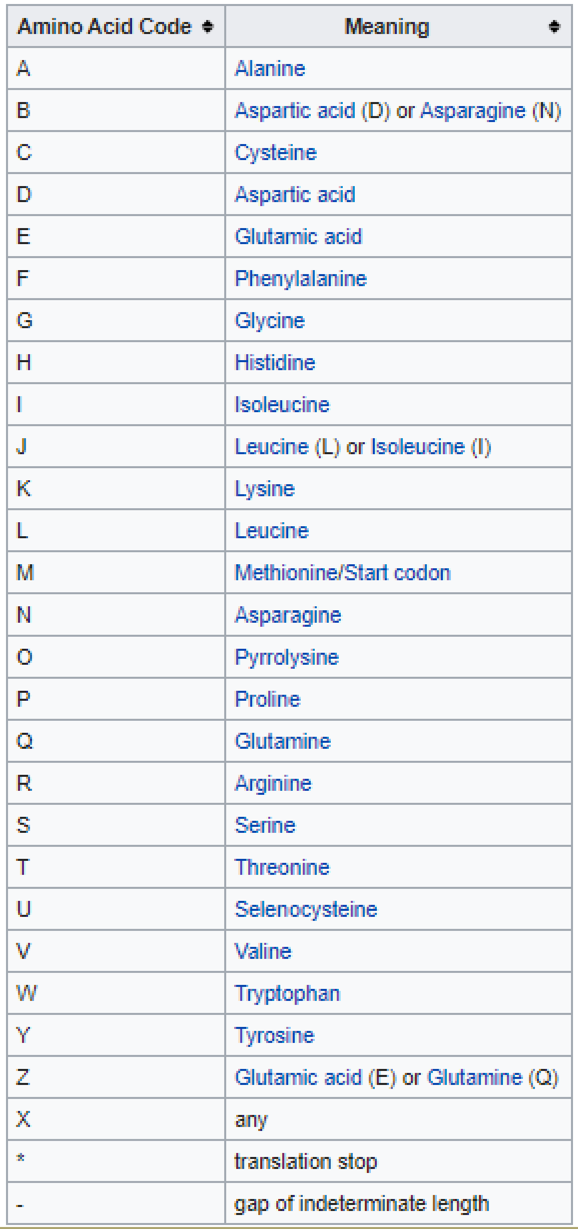
Some of the databases follow the FASTA format. FASTA is a text-based format for representing either nucleotide sequences (ex: DNA, RNA) or amino acid (protein) sequences using single-letter codes like so:

Table 2. Nucleic Acid representation in FASTA



Nucleic Acid Code	Meaning	Mnemonic
A	A	**A**denine
C	C	**C**ytosine
G	G	**G**uanine
T	T	**T**hymine
U	U	**U**racil
R	A or G	pu**R**ine
Y	C, T or U	p**Y**rimidines
K	G, T or U	bases which are **K**etones
M	A or C	bases with a**M**ino groups
S	C or G	**S**trong interaction
W	A, T or U	**W**eak interaction
B	not A (i.e. C, G, T or U)	**B** comes after A
D	not C (i.e. A, G, T or U)	**D** comes after C
H	not G (i.e., A, C, T or U)	**H** comes after G
V	neither T nor U (i.e. A, C or G)	**V** comes after U
N	A C G T U	**N**ucleic acid
-	gap of indeterminate length	

Table 1. Amino Acid representation in FASTA



Amino Acid Code	Meaning
A	Alanine
B	Aspartic acid (D) or Asparagine (N)
C	Cysteine
D	Aspartic acid
E	Glutamic acid
F	Phenylalanine
G	Glycine
H	Histidine
I	Isoleucine
J	Leucine (L) or Isoleucine (I)
K	Lysine
L	Leucine
M	Methionine/Start codon
N	Asparagine
O	Pyrrolysine
P	Proline
Q	Glutamine
R	Arginine
S	Serine
T	Threonine
U	Selenocysteine
V	Valine
W	Tryptophan
Y	Tyrosine
Z	Glutamic acid (E) or Glutamine (Q)
X	any
*	translation stop
-	gap of indeterminate length

Source: https://en.wikipedia.org/wiki/FASTA_format

The FASTA format also includes comments and identifying information for that sequence in a 'header' that starts with a ">" character. For example, a protein sequence could be shown as:

```
>SEQUENCE_1
MTEITAAMVKELRESTGAGMMDCKNALSETNGDFDKAVQLLREKGLGKAAKKADRLAAEG
LVSVKVSDDFTIAAMRPSYLSYEDLDMTFVENEYKALVAELEKENEERRRLKDPNKPEHK
IPQFASRKQLSDAILKEAEEKIKEELKAQGKPEKIWDNIIPGKMNSFIADNSQLDSKLTL
MGQFYVMDDKKTVEQVIAEKEKEFGGKIKIVEFICFEVGEGLEKKTEDFAAEVAAQL
```

Source: https://en.wikipedia.org/wiki/FASTA_format

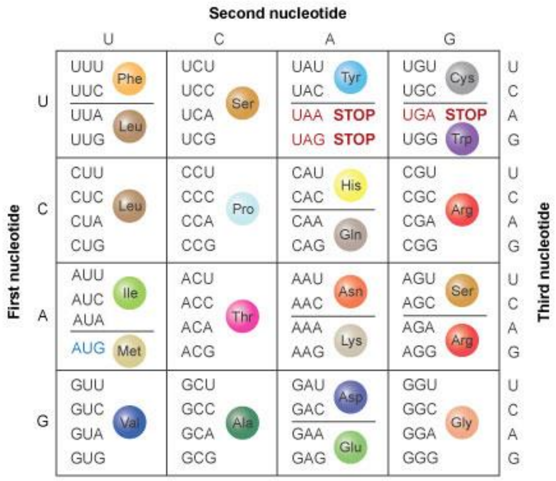
For this project, I looked at all the coding DNA sequnces available in Ensembl (DNA sequences that code for proteins) for *Homo sapiens* (humans).

I wrote functions that:

- print out the first 2 lines to see what a header + sequence looks like in the file
- find the CDS at a given chromosomal location
- print a list of all the genes in the *Homo sapiens* CDS database

Expansions:

- count the number of nucleotides (A, C, T, G respectively) in a given sequence
- translate CDS sequences into amino acid sequences (https://www.nature.com/scitable/topicpage/translation-dna-to-mrna-to-protein-393/)



> 3 nucleotides codes for 1 amino acid

> Each amino acid is represented by one letter in FASTA (see table 2)

DNA sequence	Pre-Sequence (replace T with U)	Amino acid sequence	FASTA amino acid sequence (table 2)
GAAATAGTA	GAAAUAGUA	Glu-Ile-Val	EIV

- compare the sequences for the same gene in 2 different organisms (use 2 different CDS files)
  - how many of the nucleotides are in the same position? How many are different?