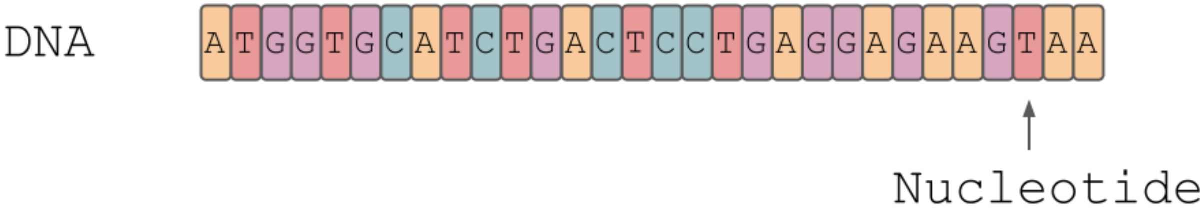
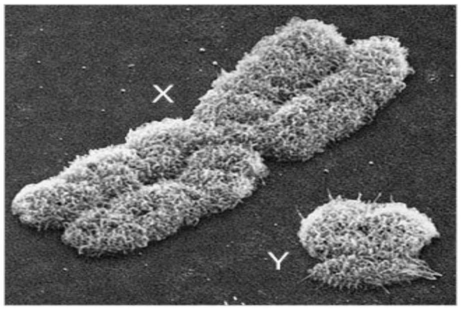
# INF115 Compulsory Exercise 1
# The Genetic Code

Working with databases, there is always some field of application, and you will have to apply it to varying subjects. In this assignment, you are fictionally employed to create a database for a customers genetic (DNA & protein) data.
Answer the following questions using SQL. Submit the code that you used to answer each question in a text document (allowed files: .sql and .txt) to mitt-uib INF115,
**<u>the deadline for submissions is 14:00 on 1st of March 2021</u>**.
Good luck!

**Background from employer:**
DNA is molecule that is essential for all living things, it provides the instructions that all cells and organisms require in order to grow, live and reproduce. Despite producing this complexity DNA is composed of strings of just 4 nucleotides, cytosine (C), guanine (G), adenine (A), and thymine (T).



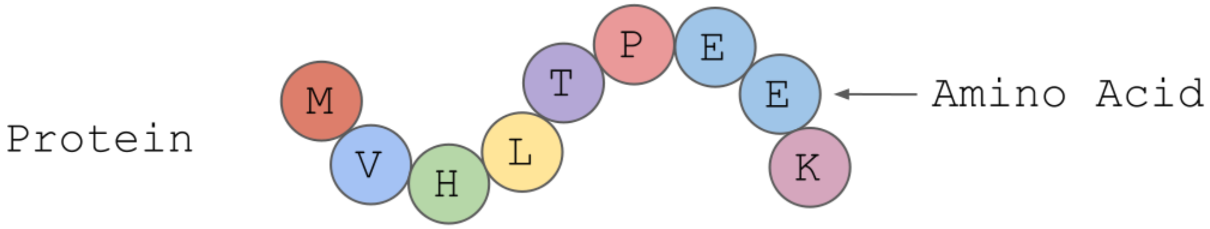
In humans DNA is structured into 23 pairs of linear strings (or chromosomes), each composed of combinations of just these 4 nucleotides, this is what we call a genome. Almost every cell in your body has it’s own copy of this genome, and each genome contains around 6,000,000,000 nucleotides. If you were to unraveled all the DNA molecules in your body and placed them end to end, it would stretch to the Sun and back several times.



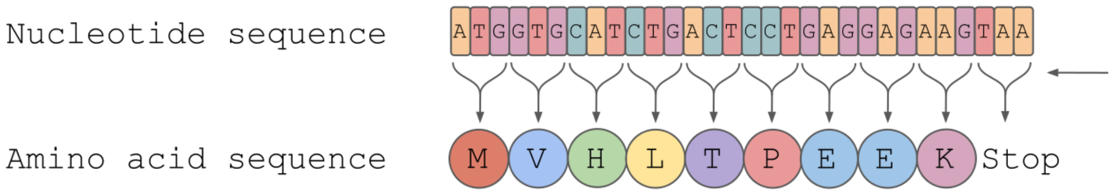
*Real electron microscopy photo of human x and y chromosomes. See how compact they are, and how much information X can contain compared to Y.*

One of the main reasons why we have DNA, is to produce proteins. Proteins are one of the main constituents of living organisms, they make up about 20% of the human body (which is quite considerable if you consider that water alone accounts for another 60%).

Each amino acid can be thought of as a building block, and there are 20 different amino acid building blocks that proteins can be made from.



DNA provides the ‘blueprints’ and ‘recipes’ for all proteins. This leads to an interesting question about how the 4 letter (A,C,G and T) nucleotide alphabet is able to encode the 20 different amino acid’s that proteins can be made from. The encoding is achieved by groups of three nucleotides together called triplets, providing the instructions for a particular amino acid.



Each of these three-letter ‘words’, formed from sequences of three nucleotides is called a codon. This triplicate encoding give us a total of 4^3=64 different codons. Because there are only 20 commonly occurring amino acids, each is encoded by one or more specific codons. There are also a small number of codons which encode a stop signal, rather than coding for a particular amino acid. These tell the biological machinery where an amino acid string should end. The DNA sequences that produce proteins usually start with the codon for the amino acid Methionine, and finish with a stop codon.

We now have all the information we need to make a database, the sql code to create some of the tables are included on mittuib from your fictional employer.

The following tables contains information on the standard genetic code:

**Codons:** (Codon_id, Codon_sequence, Position1*, Position2*, Position3*, Amino_acid_id*)
**Amino_acid_nomenclature:** (Amino_id, Symbol, Name*, Code)
**Amino_acid_properties:** (Name, Molecular_mass, Polarity, Charge)
**Nucleotides:** (Symbol, Name, Type)

Primary keys are underlined. Foreign keys are marked with *'s.

**Codons:** This table describes which 3 letter nucleotide 'word' cord for which amino acid.
Codon_id: A unique identifier for each of the 64 possible codons. Primary key.
Codon_sequence: The 3 letter DNA sequence of the codon.
Position1: The 1st Nucleotide of the codon. Foreign key to Symbol in the Nucleotides table.
Position2: The 2nd Nucleotide of the codon. Foreign key to Symbol in the Nucleotides table.
Position3: The 3rd Nucleotide of the codon. Foreign key to Symbol in the Nucleotides table.
Amino_acid_id: A foreign key to Amino_id in the Amino_acid_nomenclature table.

**Amino_acid_nomenclature:** This tables describes the naming conventions for amino acids.
Amino_id: A unique identifier for each of the amino acids and stop signals. Primary Key.
Symbol: A single letter identifier for each amino acid.
Name: The full name of each amino acid. Foreign key to Name in the Amino_acid_properties table.
Code: The 3 letter abbreviation of each amino acid.

**Amino_acid_properties:** This table describes the properties of the individual amino acids, which contribute to the overall function of the protein.
Name: The the full name of each amino acid. Primary Key.
Molecular_mass: The molecular weight of each amino acid.
Polarity: A description of if the overall polarity of the amino acid.
Charge: A description of if the overall charge of the of the amino acid.

**Nucleotides:** This table contains information about the nucleotides that make up the 3 letter codons.
Symbol: The one letter symbol for each nucleotide Primary Key.
Name: The full name of each nucleotide.
Type: The class of molecule that each nucleotide belongs to.

Answer the following questions using SQL. Submit the SQL code that you used to answer each question in a text document to mitt-uib INF115. The code for creating the **Codons, Amino_acid_properties** and **Nucleotides** tables and data are available on Mitt-uib (note that the foreign key from **Codons** to **Amino_acid_nomenclature** has not been included at this stage). The overall contribution of each section to the final marks for this assignment is listed in the brackets after each section header.

## 1. Single table queries (25%)

a) Write a query to count the number of codons.

b) Display all of the positively charged amino acids with a mass greater than 150.

c) Show all of the nucleotides of the "Purine" type, sorted alphabetically by nucleotide symbol.

d) Select all Codon_sequences that have the same nucleotide in positions 2 and 3.

e) Show the Codon_sequence and Amino_acid_id of amino acids encoded by just a single codon (For example the amino acid with id 'a11' is only encoded by the codon 'ATG').

## 2. Creating tables and modifying tables (25%)

a) Create the **Amino_acid_nomenclature** table, select data types that you think best represent the data (see table below). Include the primary key and foreign key (be careful of the null values for the stop codons in the Name field).

b) Insert the values for the **Amino_acid_nomenclature** data from the table below (at the bottom of the document), into the table in the database.

c) Add the following constraint rules to the **Amino_acid_properties** table:

i) Molecular_mass, should be greater than 70 and less than 210.

ii) Charge should be one of "uncharged' "positive" or "negative".

d) Add a foreign key to the **Codons** table referencing the amino_acid_id in the **Amino_acid_nomenclature** table.

## 3. Multiple table queries (25%)

a) List all of the codons encoding a stop signal (that do not code for an amino acid).

b) Display all of the Codon_sequence(s) that start with a nucleotide called Cytosine.

c) Write a query to return the Codon_sequence for all amino acids sorted from smallest to highest molecular mass.

d) Count the number of uncharged amino acids where the Codon_sequence ends with an “A”.

e) List the Codon_sequence and the amino acid Names for uncharged amino acids with a molecular mass between 130 and 150.

## 4. Advanced queries (25%)

a) Return a count of the number of nucleotides that are purines and the number that are pyrimidines.

b) List the Amino acid symbol for all Codon_sequences composed of just a single nucleotide (for example 'AAA', 'CCC', and so on), sort these by amino acid Name.

c) Write a query to display the Codon_sequence for all the polar amino acids with a name that finishes with 'ine', where the first nucleotide in the codon is a purine.

d) Make a count of how many of the codons would result in polar or nonpolar amino acids.

e) Further subdivide the count in 4d, by the Charge column in the **Amino_acid_properties** table (to end up with a total of 4 categories polar/uncharged, polar/positive, polar/negative, nonpolar/uncharged).

Data for Amino_acid_nomenclature (question 2.b)

| Amino_id | Symbol | Name | Code |
|---|---|---|---|
| a1 | A | Alanine | Ala |
| a2 | C | Cysteine | Cys |
| a3 | D | Aspartic acid | Asp |
| a4 | E | Glutamic acid | Glu |

| | | | |
|---|---|---|---|
| a5 | F | Phenylalanine | Phe |
| a6 | G | Glycine | Gly |
| a7 | H | Histidine | His |
| a8 | I | Isoleucine | Ile |
| a9 | K | Lysine | Lys |
| a10 | L | Leucine | Leu |
| a11 | M | Methionine | Met |
| a12 | N | Asparagine | Asn |
| a13 | P | Proline | Pro |
| a14 | Q | Glutamine | Gln |
| a15 | R | Arginine | Arg |
| a16 | S | Serine | Ser |
| a17 | T | Threonine | Thr |
| a18 | V | Valine | Val |
| a19 | W | Tryptophan | Trp |
| a20 | Y | Tyrosine | Tyr |
| a21 | | | Stop |
| a22 | | | Stop |
| a23 | | | Stop |