CSC 314, Bioinformatics Lab #5: **Name:**_____
GenBank Nucleotide Database

GenBank (http://www.ncbi.nlm.nih.gov/genbank/) is a genetic sequence database hosted by the National Center for Biotechnology Information (NCBI) under the National Library of Medicine (NLM) at the National Institutes of Health (NIH). Genbank is a collection of publicly available DNA sequences and is part of the International Nucleotide Sequence Database Collaboration, which also includes the DNA DataBank of Japan (DDBJ) and the European Molecular Biology Laboratory (EMBL). The three organizations exchange data on a daily basis and therefore have identical sequence records (but different interfaces and formats).

GenBank records can be accessed in three ways:

1. Searching GenBank for sequence identifiers (accession numbers) or by annotations (such as keywords).
2. Searching GenBank for a similar sequences to a query sequence using BLAST (we will discuss this in detail at a later date)
3. Searching GenBank and downloading sequences programmatically (we will discuss this in more detail at a later date)

This lab is designed to give provide a tutorial of the GenBank database. During this lab, you will answer questions designed to walk you through understanding and using GenBank to find information about genetic sequences.

Part I. GenBank Statistics.

From the GenBank homepage (https://www.ncbi.nlm.nih.gov/genbank/), click on the GenBank menu on the top left and select *Statistics*. Note the exponential growth of GenBank as illustrated in the two graphs at the top of the page. For the questions that follow, use the *GenBank* statistics, rather than the Whole Genome Shotgun (WGS) statistics.

Complete the table below to indicate the number of bases and sequences in GenBank at various times. Round your answers to the nearest units that are indicated. For example, if the answer is in thousands and the value was 23,687, your answer should be 24 thousand. The following numbers are given as reference:
1,000,000 (1 million); 1,000,000,000 (1 billion).

Time	Number of Bases	Number of Sequences
December 1982 (GenBank is first developed)	_____ thousand	_____ hundred
December 2001 (1st draft of human genome completed)	_____ billion	_____ million
February 2019 (most recent release)	_____ billion	_____ million

Part II. Searching

There are 3 aspects related to searching: basic searching, searching using limits (filters), and advanced searching by field. ***Note: GenBank will remember your filters in future searches unless they are explicitly cleared.***

1. Enter HBB into the search box at the top of the page. Note that when searching GenBank, the selected search type next to the search box should say *Nucleotide*. Press enter to carry out the search. (Note that HBB is the gene name for hemoglobin beta, which codes for the beta unit of hemoglobin). All of these databases contain nucleotides, and so the total number of nucleotide sequences found is also returned. How many sequences are found?

2. The above is a keyword search, and so not all entries correspond to the HBB gene (for example, if HBB is in the description, but not the name, the record will be returned). Click on Advanced, change the field to Gene Name, and search again for HBB. How many sequences from the *Nucleotide* database are found?

3. On the right hand side of the screen, click on *Homo sapiens* (humans), which filters the results to only include records for *Homo sapiens*. How many sequences in the nucleotide database are there for the HBB gene in humans?

4. Look on the left-hand side of the screen under *Molecule types*. How many entries correspond to mRNA molecules?

5. On the left-hand side of the screen, click on *Custom Range* under *Release Date*. How many entries of the human HBB gene of any molecule type were published since 1/1/2015. Note: make sure to clear any filters that are no longer relevant.

Part III. The Data

A sample GenBank entry is given here: https://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html. The links on this page will take you detailed descriptions of each element or field.

Each entry can be divided into three parts:

1. the ***header*** contains summary information about the entry and references to scientific literature. For example, the first line includes the # of base pairs, the molecule (DNA in this case), and the last modification date.
2. the ***feature table*** (starting with the word Features) contains information about the sequence along with the position (i.e., nucleotide positions) of the corresponding feature, with 1 being the position of the first nucleotide in the sequence.
3. the ***sequence*** contains the sequence (duh), with the number on the left corresponding to the nucleotide position of the first nucleotide in that row. Nucleotides are displayed in groups of 10 for ease of counting.

From the sample page above, click on the links for the terms below and briefly describe the following fields:

1. ACCESSION:

2. CDS

3. gene

4. translation

5. complement

Part IV. Analysis of the "violence gene" Monoamine oxidase A (MAOA)

For more information about this gene, see: http://www.bbc.com/news/science-environment-29760212

1. How many human (*Homo sapiens*) RefSeq entries are there for the gene MAOA in GenBank? Hint: make sure to search specifically for the *gene name* and for *Homo sapiens*, based on what you learned previously. The **RefSeq** (Reference Sequence) collection is a comprehensive, integrated, non-redundant, well-annotated set of sequences, including genomic DNA, transcripts, and proteins.

Open the entry with Accession Number NG_008957.2 to answer the remaining questions.

2. When was this entry last modified?

3. What chromosome is this gene on (Hint: this can be found by looking at the *source* feature)?

4. The MAOA gene spans what positions of this DNA sequence?

5. What are the first 5 nucleotides of the gene? (Hint: click on the *gene* feature)

6. This gene contains how many (inferred) exons? (Hint: you have to look at the exon features which are numbered)

7. Click on the CDS feature (note that since the CDS covers multiple exons, it will span multiple disjoint regions of the DNA sequence). What are the last 3 codons of this protein, and what do they code for (Note: it is recommended that you use the table in your notes or the Expasy Translate tool (https://web.expasy.org/translate/) to translate these)?

8. Click on the PubMed link for the article "Abnormal behavior…", which is in the *header* of the GenBank entry, under Reference 2 (Note that the link is below the article title). What type of mutation (nonsense, frameshift, missense, or silent) was found to be associated with abnormal behavior? What exon is this mutation found in?