CSC 360, 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 annotations (such as keywords).
2. Searching for a particular sequence or similar sequences using BLAST (we will discuss this in detail at a later date)
3. Searching and downloading sequences programmatically (we will discuss this in more detail at a later date)

This lab is designed to give you 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.

A. How many sequences and number of bases were in GenBank

1. in December of 1982, when GenBank was first developed?

2. in December of 2001, when the first draft of the human genome was published?

3. in the most recent release, in February of 2017?

B. How many sequences have been added to GenBank in the past year, from February 2016 to February 2017? Round your answer to the nearest million.

C. How many bases have been added to GenBank in the past year, from February 2016 to February 2017? Round your answer to the nearest billion.

Part II. Searching

There are 3 aspects of 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). The search is carried out across 3 databases: Genbank (the 'Nucleotide' results), the EST database (the 'EST' results), and the GSS database (the 'GSS' results). GSS are Genome Survey Sequences similar to ESTs but based on DNA rather than mRNA. All of these databases contain nucleotides, and so the total number of nucleotide sequences found is also returned. How many total nucleotide 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). Click on Advanced, change the field to Gene Name, and search again for HBB. How many total nucleotide sequences 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 total nucleotide sequences are there for the HBB gene in humans?

4. Look on the left-hand side of the screen. 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 IV. 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 to a detailed description 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 corresponding position (i.e., nucleotide positions, 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 the row. Nucleotides are displayed in groups of 10 for ease of counting.

Briefly describe the following fields:

1. ACCESSION:

2. CDS

3. gene

4. translation

5. complement

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

Reference: 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: look under 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 exons? (Hint: you have to look at the exon features)

7. Click on the CDS feature. What are the last 3 codons of this protein, and what do they code for? (Use the table in the notes to translate these, and then confirm by looking at the translation given by GenBank)

8. Click on the PubMed link for the article "Abnormal behavior…". What type of mutation (nonsense, frameshift, missense, or silent) was found to be associated with abnormal behavior? What exon is this mutation found in?