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. Searching Tutorial***

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 subunit of hemoglobin). The total number of nucleotide sequences found is returned. The search results are shown in the screenshot below. A total of 7,763 results are returned. There may be information above this. However, note that these are references to other databases and are not what we are interested in now.

A screenshot of a computer

Description automatically generated

1. The above search will carry out a search for “HBB” in all fields of a record (such as title, description, etc). Therefore, not all entries will correspond to the HBB gene specifically (for example, if HBB is in the description). To limit the search to a gene, click on Advanced, change the field to Gene Name, and search again for HBB. You should find 4,183 nucleotide sequences. Note this carries out a search for HBB [Gene Name] – the brackets provide a shortcut for searching by a particular field.
2. On the right-hand side of the screen, click on *Homo sapiens* (humans), which filters the results to only include records for *Homo sapiens*. Confirm that there are 1,338 sequences in the nucleotide database for the HBB gene in humans.
3. Look on the left-hand side of the screen under *Molecule types*. Confirm that there are 13 mRNA entries for Homo sapiens.
4. On the left-hand side of the screen, click on *Custom Range* under *Release Date*. And enter the from date corresponding to 1/1/2020. Carry out the appropriate search to confirm that 400 nucleotide sequences for the HBB gene (in any organism) have been published since 1/1/2020. Note: you will need to clear any filters that are no longer relevant (by selecting “Clear all” under the filter panel on the left side of the screen, or by updating the search box).

***Part II. 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 GenBank 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 of an entry includes the number 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 at the bottom of the entry.
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 navigating.

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 III. 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. The number of RefSeq entries appears on the left-hand side of the page under Source Databases.

Open the entry with Accession Number NG\_008957.2 to answer the remaining questions.

1. When was this entry last modified?
2. What chromosome is this gene on (Hint: this can be found by looking at the *source* feature)?
3. The MAOA gene spans what nucleotide positions of this DNA sequence (Hint: look at the *gene* feature)?
4. What are the first 5 nucleotides of the gene? (Hint: click on the *gene* feature)
5. This gene contains how many (inferred) exons? (Hint: you have to look at the exon features which are numbered)
6. 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, using the codon table from your notes. Also specify the 1 letter amino acid codes, which can be found at this link: <https://molbiol-tools.ca/Amino_acid_abbreviations.htm>. Does this match the end of the translated sequence found in the feature table? Why or why not?
7. Click on the PubMed link for the article "Abnormal behavior…", which has a PubMed ID of 8211186,

which is in the *header* of the GenBank entry, under Reference 2 (Note that the PubMed 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?