CSC 314, Bioinformatics Lab #6: **Name:**_____
GenPept Protein Database

GenPept (http://www.ncbi.nlm.nih.gov/protein) is a protein 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). GenPept is a collection of publicly available protein sequences from several sources, including SwissProt, PIR, and PRF, as well as translated coding regions from GenBank. Because GenPept and GenBank are both hosted by NCBI, they can be searched in a similar fashion and their formats are similar.

This lab is designed to give you a tutorial of the GenPept database. During this lab, you will answer questions that will walk you through using GenPept to find information about proteins.

Part I. Searching and HBB entry example

As with GenBank, GenPept allows for *basic* and *advanced* searching. Once a search is carried out, filters can be applied by clicking on the appropriate link on the left or right hand side of the screen. Go to http://www.ncbi.nlm.nih.gov/protein/ to begin. ***Note: GenPept will remember your filters in future searches unless they are explicitly cleared.***

1. Enter HBB into the search box, and press enter. (Recall that HBB is the gene name for hemoglobin beta). How many protein sequences are found?

2. This is a keyword search, and so not all entries correspond to the HBB protein (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. This searches for all proteins corresponding to the gene name HBB. How many total sequences are found?

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

4. How many of these sequences are based on GenBank nucleotide sequences? Note: Click on GenBank under Source Databases on the left hand side. If GenBank is not visible, then click on Customize and select GenBank. You will then need to click the 'search' button to refresh the results.

5. Click on the GenBank link again to remove this filter. Then click on RefSeq. Recall that the **RefSeq** (Reference Sequence) collection is a comprehensive, integrated, non-redundant, well-annotated set of sequences, including genomic DNA, transcripts, and proteins. Because there is only one RefSeq entry, you will be taken directly to this entry. What is the accession number of this entry, and when was it last modified?

6. What is the length of this protein?

7. What amino acid positions correspond to the Hb-Beta like region?

8. The main role of hemoglobin is to transport oxygen. In order to do this, hemoglobin must bind with an iron-containing molecule known as a heme group. How many amino acids are involved in heme binding?

9. What are the first five amino acids involved in heme binding (you can give the one letter codes)?

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

1. How many human (*Homo sapiens*) RefSeq protein entries are there for the gene MAOA in GenPept?

2. Note that different *isoforms* refer to different forms of the same protein, either due to alternative alleles or from alternative splicing. What is the length of isoform 1 and what is the length of isoform 2?

3. Look at the entry of for isoform 2. Why is this isoform shorter than the other isoform? (Look at the transcript variant description in the comment section.) Note: The UTR is short for "untranslated regions" which refers to a region of DNA that is transcribed into mRNA but not translated into protein.

4. There are two functional regions in this protein, one of which is the monoamine oxidase region. A monoamine oxidase is an enzyme that breaks down neurotransmitters such as serotonin and dopamine. What is the location (the amino acid positions) of the monoamine oxidase region in this sequence?

5. Mutations in *MAOA* can result in Brunner syndrome (you should find the corresponding entry at http://www.omim.org)>

    a. Look under clinical synopsis, and identify the two features associated with the central nervous system.

    b. What is the mode of inheritance of this disorder?

    c. The entry states that "In a boy with Brunner syndrome…Piton et al. (2014) identified a hemizygous mutation in the MAOA gene…which was found by high-throughput sequencing of coding exons of intellectual disability genes…[and the mutation] was also present in the [boy's] unaffected mother." Although both the boy and his mother had a copy of mutated gene, why did the boy have Brunner's syndrome while his mother did not? What is meant by "hemizygous mutation" in this context?