CSC 314**, Bioinformatics Lab #6: Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**GenPept Protein Database (+Additional Tools)**

The NCBI Protein Database (<http://www.ncbi.nlm.nih.gov/protein>) is hosted by the National Center for Biotechnology Information (NCBI) under the National Library of Medicine (NLM) at the National Institutes of Health (NIH). This database 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 the Protein Database 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 provide you with a tutorial of the NCBI Protein database. During this lab, you will answer questions that will walk you through using the database to find information about proteins.

***Part I. Searching and HBA1 entry example***

As with GenBank, the Protein Database 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: your filters will be remembered in future searches unless they are explicitly cleared.***

1. Enter HBA1 into the search box, and press enter. (Note that HBA1 is the gene name for one of the hemoglobin alpha subunits). How many protein sequences are found?
2. This is a keyword search, and so not all entries correspond to the HBA1 protein (for example, if HBA1 is in the description, but not the name). Click on Advanced, change the field to Gene Name, and search again for HBA1. This searches for all proteins corresponding to the gene name HBA1. 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. 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. There should be one RefSeq record, which has accession number NP\_000549.1. Open this record. When was this entry last modified?
5. What is the length of this protein (how many amino acids)?
6. What amino acid positions correspond to the Hb-alpha like region (Hint: this is a *region* feature)?
7. 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, and what are the first five amino acids involved in heme binding (you can give the one letter codes)? Hint: look at the *Site* feature with a note for “heme binding site [chemical binding]”. Note that the amino acids are dispersed across the sequence, but are close to each other when the protein is folded in its proper 3D structure).

***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 the NCBI Protein Database?
2. Note that different *isoforms* refer to different versions of the same protein, either due to alternative alleles or alternative splicing that produces different CDS regions of the mRNA. What is the length of isoform 1 and what is the length of isoform 2?
3. Look at the entry of for isoform 2, with accession number NP\_001257387.1. 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 (look at the *Brunner Syndrome; BRNRS* entry at <http://www.omim.org>)
   1. Look under clinical synopsis, and identify the last three features associated with Behavioral Psychiatric Manifestations
   2. What is the mode of inheritance of this disorder?
   3. In the second paragraph under the Molecular Genetics section, 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." Note that an individual who is hemizygous only has one copy of the gene. Although both the boy and his mother had a copy of the mutated gene, why did the boy have Brunner's syndrome while his mother did not? Why did the boy have one copy of the gene, but his mother had two?

***Part III. UCSC Table Browser***

1. Use the UCSC Table Browser to get the CDS exons for MAOA, using the settings below. Display the output using one FASTA record per gene.

**A screenshot of a computer

Description automatically generated**

* 1. Specify the FASTA header for each result (the header is the line that begins with a > sign)
  2. For each CDS, specify the first and last codons (these should be the same for each entry). Do these codons make sense? Why or why not?

***Part IV. Gene Database***

1. According to the *Gene* entry for human MAOA
   1. What tissue has the highest expression?
   2. What is the mouse ortholog for the gene MAOA?
2. The gene TP53 is one of the most important genes associated with cancer in humans. According to the *Gene* entry for human TP53
   1. What three domains are in the protein?
   2. What is the mouse ortholog for TP53?