Bioinformatics, computer lab II
University of Potsdam WS 2019/2020, S. Hartmann

Online resources: biological sequences

You will explore different biological databases by looking up information about the same globin gene/protein. The purpose of this exercise is to get familiar with the different interfaces and search functions of these databases. Of course, you'll learn more about the globin family as well.

1 Primary DNA sequence databases: GenBank

Look up the beta globin DNA sequence from human (*Homo sapiens*) in GenBank, which was introduced as one of the three primary sequence databases and is available at NCBI: `http://www.ncbi.nlm.nih.gov`. Select the "Nucleotide" database and search for the sequence ID 'NM_000518'; the underscore is part of the ID and needs to be included.

☞ What are the accession number and other IDs or names that are associated with this sequence? Has the entry been revised yet?

☞ What is the length of the DNA sequence?

☞ Does the entry include information about the function of the protein?

☞ Are there any publications associated with this entry?

☞ As discussed in lecture, the line between primary and secondary databases can be blurred. In the DNA entry, there is a line starting with the word 'COMMENT'. It explains that this entry is a "reviewed refseq". What does this mean?

☞ Can you download the DNA sequence? If so, which formats are available?

☞ Can you access and download the corresponding protein sequence? Does the protein sequence have the same ID as the DNA sequence? Which formats are available? (hint: search for 'protein_id' within the features and follow the link to the entry for the corresponding protein sequence.)

2 A database of annotated protein sequences: UniProt

Look up the same sequence in Uniprot at `http://www.uniprot.org`: Does an entry with the ID from the DNA sequence 'NM_000518' or the protein entry (NP_000509) exist in UniProt?

On the results page, click on the entry with the accession P68871 and the name HBB_HUMAN. Briefly browse through the page to see what kind of information is available for this entry.

☞ What kind of information do you find at the UniProt page that was not available in the GenBank entry you just visited?

☞ What accessions / names / IDs are associated with the entry? Are any of them the same as you found in GenBank?

☞ Does the entry tell you what the accession numbers of this sequence is in a primary database (i.e., NCBI, EMBL, DDBJ)? Do you find links to the entries of this sequence in NCBI, EMBL, DDBJ?

☞ Has this sequence been 'reviewed' yet? What does that mean?

☞ Can you download the sequence? In which formats?

☞ Which cross-references to secondary databases are available?

☞ Are there sequences in UniProt that have 100% sequence identity to the human hemoglobin sequence? How many, and from which organisms?

☞ Is a solved 3D structure available for this protein sequence? One or more?

3 OPTIONAL: Structure databases

3.1 PDB

The Protein Structure Database can be accessed at `http://www.rcsb.org/`. Again, you can use the ID from another database to search for and find structures, but the PDB uses different IDs. Visit the PDB site and search for the Uniprot ID P68871, the sequence for human beta-hemoglobin. How many solved structures are available at PDB that are associated with this UniProt ID?

☞Click on any one of them, but remember (or write down) the ID of the one you click on.

You'll be taken to summary information of the molecule and its structure, indicated by the tab "Structure Summary", highlighted in dark blue. Other tabs with more detailed information are to the right of it.

☞Just by looking at the titles of these tabs – which ones do you think contains primary data (i.e., directly submitted by the researcher who solved the structure), and which contain secondary data (additional information and analysis results)?

☞ Explore any other information on the PDB sites that interests you.

3.2 CATH

Visit the CATH website at `http://www.cathdb.info/`. Where does the data in CATH come from? (Hint: On the top of the page, click on "About")

Globins are in the CATH Superfamily "1.10.490.10", search for this ID and click on it on the results page. You will be taken to a summary of this superfamily. A lot of information is available here!

☞ Click on "Classification / Domains" (on the left, under "Superfamily links"). To which class, architecture, and topology do the domains in the globins superfamily belong? What

does this mean?

☞ Explain in your own words what the graphic under the heading "CATH Domains" represents.

☞ Explore any other information on the CATH sites that interests you.

4 Summary

☞ Briefly summarize the information that can be looked up about the human beta-globin on the different resources. Try to generalize as much as possible.