In this exercise we will explore a few important biological databases. First, we'll take a look at the NCBI sequences database. Then, we'll learn about the Uniprot protein sequences database. And finally we'll see how to search in the RCSB- Protein data bank for 3D protein data.

Let's start with the NCBI site. We will use the built-in search engine **ENTREZ** to search through the NCBI databases. We can search for:

-Nucleotides and proteins databases.

- Articles (PubMed).

- OMIM (online Mendelian Inheritance in Man).

dbSNP and Clinvar for known human variations

And even for additional databases…

Note that at this stage we are performing a text search, by searching for key words, terms, gene names, etc. In the next session we'll talk about Blast, which can search sequence databases by using sequence similarity methods within the sequences themselves.

For now, enter the NCBI website through the following link: www.ncbi.nlm.nih.gov

We will first perform a search through the NCBI "All Databases". We will search for all results where the "vascular endothelial growth factor" words appears. Please note that there are more efficient ways to use the search, but right now we'll use the most straightforward way simply to become familiar with the database.

Write " vascular endothelial growth factor" in the search box, make sure that "All Databases" is selected under the left scroll bar and click on search.

The entrez page will upload and present all existing databases with the number of relevant hits marked. Three important and commonly used databases are the Nucleotides, Proteins and Genes databases. We will focus on the results of the Nucleotide sequence database, so click on its name. Alternatively, we could also enter directly to its results page by performing the same query straightaway in the "Nucleotide" database.

As you see, search results are displayed in a summary format. Results are initially displayed with 20 items per page. You can change the display default properties by using the Display Settings menu.

Click on the "homo sapiens" link that appears on the right to select all sequences originated from humans. Each sequence result includes two fields: an accession number and a GI.

An accession number is the unique identifier for a sequence record. An accession number applies to the complete record and is usually a combination of letters and numbers. This accession number can be used in many databases and bioinformatics tools.

Accession numbers do not change, even if information in the record is revised at the author's request. If there is any change to the sequence data (even a single base), the version number will be increased, for example version 1 will be changed to version 2 by adding point 2 (.2) at the end of the accession number.

There are many types of accession numbers, depending on which company/database they originated from. For example: Refseq, Genbank, Swissport databases, and others as well. One known common used database is **RefSeq**. Records of reference sequences from the **RefSeq** database have a distinct accession number format that begins with two letters followed by an underscore bar and six or more digits, for example:

NT_123456 for constructed genomic contigs

NM_123456 for mRNAs

NP_123456 for proteins and

NC_123456 for chromosomes

GI - "GenInfo Identifier" is a sequence identification number which is constructed from numbers. If a sequence changes in any way, a new GI number will be assigned.

The accession.version system of sequence identifiers runs in parallel to the GI number system. When any change is made to a sequence, it receives a new GI number AND an increase to its version number (for example U12345**.1**, 123456 → U12345**.2**, 7891011).

To conclude, one gene can have many different accession numbers derived from different databases. In addition, duplicate information also exists in data bases for many reasons. For example, when different groups work on the same sequence and give it different accession numbers and even different names! Therefore, it is important to know how to search for the relevant updated accession numbers and how to find the different names representing a gene.

Click on the first result. Now you've entered the human VEGFB mRNA page. The entry contain various types of data about this mRNA, such as the sequence length, a date of the last update and a list of publications related to this transcript. The sequence itself will appear at the end of the page in a Genbank format. There is no description line in this type of format. The sequence is presented as 60 letters per line where each 10 letters are separated by a gap. The amino acid numbering is presented at the beginning of each line. A more commonly used format is FASTA. Click on the Fasta link that appears at the head of the page. You can see that the sequence begins with a single-line description, followed by lines of sequence data. The description line must begin with a greater-than (">") symbol in the first column. Each line contains 60 letters.

Now let's move to the uniprot site which is a comprehensive proteins database.
http://www.uniprot.org/

The UniProt**KB**- UniProt Knowledge Base consists of two sections:

The first one is the "UniProtKB/TrEMBL"- for unreviewed records. It includes Computationally analyzed records that await full manual annotation.

The second one is the "UniProtKB/Swiss-Prot" for reviewed records. It includes manually annotated records with information extracted from literature and curator-evaluated computational analysis. Each record that is transferred to this database is automatically removed from the TrEMBL in order to prevent redundancy.

Search for the bcl2 protein at the search window, the results can be further filtered on the left panel, for swissprot result or tremble ones. In general, we will always prefer to work on the swissprot database unless there is no other relevant record.

Now let's Search for the hemoglobin alpha 1 protein by its accession number P69905 at the UniProtKB dataset. The result page will show details about the protein such as name and origin, description, functionality, description of known variants, secondary and tertiary structures, protein sequence in different formats, annotation on the sequence like GO, data on related diseases, motifs and more. It will also show related data from different databases such as: UCSC, KEGG, ENSEMBL and so on. There is also an option to blast the protein directly from the page.

Now let's move on to the third site – the RCSB- PDB database which includes three-dimensional protein structures. http://www.rcsb.org/pdb/home/home.do

The Protein Data Bank (PDB) archive - includes information regarding the 3D shape of proteins, nucleic acids, and complex assemblies. This type of information helps researchers to better understand the function of the proteins, which may have implications on health and related diseases.

The RCSB portal includes additional data on the macromolecule from different databases (like: SCOP or Uniprot) and it enables presenting the structure online, presenting the sequence in FASTA format and performing other analysis on the structure. Currently, there are about 150,000 available biological macromolecular structures from different organisms (92% are proteins), most of them solved by the X-RAY technique.

Let us search for the human hemoglobin protein structure. Enter its PDB accession number 4HHB to the search window. You can also search by protein sequence, author name, ligand and more.

The data on the hemoglobin protein is presented in different tabs. The "Structure Summary" tab includes data such as the author name, organism, expression system that was used, the method and resolution, and so on. We can see that this human protein was solved using the X-RAY DIFFRACTION method with a resolution of 1.74 Å (angstrom). We consider structures that were solved with resolution lower than 3 Angstrom to be of good quality.

In the" Sequence" worksheet we can find a graphic view of the sequence including the secondary structure presenting helixes, beta turn and loops among others structures.

The structure can be visualized through the “3D view” tab. Left click of the mouse will drag and rotate it. Scrolling the mouse will zoom in and out the structure. Placing the mouse over it will identify atoms and bonds. For more presentation option use the menu at the bottom.

You can download the PDB file through the “Download file” link that appears at right. Click on it and choose “PDB format”. The PDB file includes the coordinates of the structure and additional data, so you can open it with a structure viewer such as the sPDBviewer to continue analyzing it.