Searching databases at NCBI

You will use Search NCBI, which is the interface for rapid interrogation of all the NCBI databases. You will explore how to efficiently query a sequence database by progressing in difficulty.
This screnshot indicates the features we will use. It is advisable to keep it open so that you can refer to it.

Reading a nucleic sequence entry

You first arrive on the home page allowing you to query all the database at once. When a query is run, the number of entries for each databases appears next to the database name. We are going to search for the entry in the database "Nucleotide" which has the accession number `Y10810`. To do this, you will enter the words to search for (here the accession number) in the query input field.

> **Question 1**
>
> How many entries can be found in the Nucleotide database ?
> Are there other databases with (a) match(es)?

Go to the Nucleotide database entry to answer the following questions (you must click on the name of the database and then on the entry number):

> **Question 2**
>
> Is it a DNA or a RNA sequence?
> From which species does the sequence in the entry come?
> Is it an eukaryote or a prokaryote?
> What is the size of the sequence contained in the entry?
> How many coding sequences (CDS) are present in this entry?
> Does this seem normal to you given the molecular nature of the sequence (DNA/RNA) and its origin (eukaryote/prokaryote)?

The first coding sequence is annotated as a uORF. It is an element that is involved in the regulation of translation of mRNAs. This is why there are 2 CDS on a mRNA.
For a more precise description of what a uORF is, you can do a search in the database Books which contains a set of books whose text is available on the web. You just have to enter `uORF` in the input field after having chosen the database *Books*.

> **Question 3**
>
> What does the 'u' in uORF stand for?
> How does a uORF function?

Let's now go back to the entry of our mRNA in the Nucleotide database to determine the peptide encoded by uORF.

Question 4

What is the position of the uORf within the mRNA?
Consult the sequence of the entry to determine the sequence of the uORF.
With the help of the genetic code translate the sequence of the uORF.
Check your translation using the protein sequence given in the entry (ligne *translation*).

Change display format and file manipulation

At the top left of the page there are different options to change the display format. Here you can find *FASTA*, *Graphics* and the menu *Display Settings*. The default format of an entry is *GenBank*. An other menu, *Send*, which is located on the right of the page, allows you to save the data either in a file (*File*), or in the temporary memory (*Clipboard*), or permanently on your MyNCBI account if you have created it.

Change the display format to *Graphics*. Look at the CDS `CPRF4b` (green), and the associated protein (red), as well as their respective information.

Question 5

What are the lengths of the CDS and the protein produced?
How many protein domains/motifs (among the elements in black) are described on this protein? What are they for?
Is there any other information? Which ones?

Switch back from display format *Graphics* to *GenBank* format.
Then save the complete GenBank entry using the *Send* → *Complete Record* + *File* in the *GenBank (full)* format, as well as in the *Fasta* format (You should obtain two files).

Question 6

What happens if you try to open files previously saved in *Fasta* or *GenBank* by double clicking ?

To read these files, you need to open any text editor, then use the menu *Fille* → *Open* to fetch the file. It is also possible :

- change the name of the file (for example by adding the extension ".txt" to the files "sequence.gb" and "sequence.fasta"),
- otherwise (if you do not have sufficient access rights, for example), save the file previously opened by the text editor by adding a ".txt" extension which will make it easier to reopen under Windows.

Note that it is always possible to copy and paste into a file, using the formats proposed in the *Display Settings* menu.

Display the sequence in the format *GenBank* using the menu *Display Settings* : you will note that it is possible to Copy and Paste this sequence if you are careful to select only the part associated with the GenBank file.

Then display the sequence in *FASTA* format then in *FASTA (text)* format.

> **Question 7**
>
> Which of these two formats is easier to copy and paste?

Targeted Query

Using the search bar

Run the query `Bacillus subtilis` in the *Nucleotide* database.

> **Question 8**
>
> Do the sequences found all come from this bacteria ?
> For records that do not come from *B. subtilis*, where does the name of this bacteria appear (consult only a few entries)?

By default, the terms of a query are searched in the whole entry. To make more relevant searches, it is necessary to specify the field in which the terms are searched. For NCBI databases, querying a specific field is done using the field name in square brackets after the search term(s)s.

Run the query `Bacillus subtilis [organism]` in the *Nucleotide* database.

> **Question 9**
>
> Do the sequences found now all come from this bacteria?

When several terms are searched for, it is possible to combine them using using the Boolean operators :

- **AND** : the two terms are both in the entries.
- **OR** : at least one of the two terms is in the input.
- **NOT** : the first term must be present in the entries and the entries that contain the second term are excluded.

> **Question 10**
>
> Guess and interpret the meaning of the additional request `Bacillus subtilis NOT Bacillus subtilis [organism]` and validate its result with the number of entries found by the two previous queries.

Via the Builder

Build queries "manually" by indicating the fields (such as `(Bacillus subtilis [organism] OR Yersinia pestis [organism]) AND MOTB [gene]`) can quickly become tedious... Fortunately, below the input field of the query, the link *Advanced* provides access to a tool for building queries and named *Builder*.

In the tool *Builder*, search in the list of available fields, the one that allows to limit the entries to that which comes from the **organism** *Bacillus subtilis*. Then, enter `Bacillus subtilis` in the input box next to the list of fields.

> **Question 11**
>
> What is the automatically created query?
> Is the automatically constructed query identical to the previous one?
> Check the consistency with the number of results obtained.

Now search for the gene sequence `MAKORIN1`, in the fish *Seriola quinqueradiata*. If you do not specify a field name for the gene name, you should get 4 entries: 2 mRNA entries (one complete and one partial) and 2 genomic sequence entries (one with the complete gene, the other with the incomplete gene).

consult these 4 entries and note behind which *qualifier* is indicated `MAKORIN1`. You can find out which field to search by consulting the list of available fields in the advanced search form.

> **Question 12**
>
> What is the query then automatically constructed by *Builder* if the fiels `[gene name]` is used for `MAKORIN1` in addition to the field `[organism]` for `Seriola quinqueradiata`?
> How many entries are then found?
> Why are the other entries lost?

Now, we will look for **proteins** with either the enzymatic function that has the EC number 5.3.1.24, or the enzymatic function 4.1.1.48, or both. It is therefore necessary to query the database *"Protein"*. To begin, find out if there is a field that matches the EC numbers. Then, answer the following questions by choosing the appropriate operators:

> **Question 13**
>
> How many entries describe a protein that has the function 4.1.1.48?
> How many entries describe a protein that has at least one of the two functions 4.1.1.48, 5.3.1.24 ?
> How many entries have the two functions 4.1.1.48 and 5.3.1.24?

History of queries

Often database query systems store the queries made by your computer since the beginning of your connection. This set of queries is called the history.

The history is accessible either in the *History* which is located on the right side of the screen, at the bottom, or by using the link *Advanced* which allows to manipulate queries). This allows you to display the results of a query again.

> **Question 14**
>
> Rerun the query that gives the gene **and** mRNA of the MAKORIN1 gene, in the fish *Seriola quinqueradiata*.

Links between databases

Search NCBI allows to query many databases (Nucleotides, Protein, PubMed, ...). Data from one database can be linked to data from another database. For example, nucleic sequences containing a gene can be linked to the proteins encoded by these genes. The links are obtained using the *Find related data* (see banner on the right of the page).

Link to the Proteins database

Question 15

From the previous query, make a link to the database *Protein* to obtain the proteins coded by the 4 entries.
How many entries do you get ?
Are the resulting protein sequences different?

In fact, you can find 2 proteins of 418 aa and 2 proteins of 435 aa because each time there is a gene and an mRNA which code the same protein. The two pairs of protein entries are generated automatically from the nucleic sequences. There is no deletion of redundant sequences which have differentaccession numbers.

Link to the Genes database

Question 16

Look for genes or mRNA that code for a protein with a function *"selenophosphate synthetase"* (with quotation marks), in the organism *"Homo Sapiens"*.
Which database should we ask?
Are the entries found redundant or correspond to different genes?

In addition to chromosomes and mRNA, there are chromosome fragments (*genomic contigs/scaffolds*) that were sequenced ans deposited in the databasee before the full genome was available. In addition, the result list contains many mRNAs resulting from alternative splicing events. Therefore, it is difficult to know how many genes are associated with this function and which mRNA corresponds to which gene.

Question 17

Follow the link to the database *Gene*. How many entries do you get?

The number of entries after the link to *Gene* is very important, because the (fragments of) chromosomes parasitize the result : indeed **all** genes contained in them are researched ...

It is possible (for example) to select a number of mRNAs (by clicking on the "check boxes" to the left of each entry)

Question 18

Now select the few mRNAs entitled "Homo sapiens selenophosphate synthetase 1 (SEPHS1)", and follow the link to the database *Gene*
How many entries do you get now? Check it out.

Filters

To avoid a tedious work of selection of the mRNAs, it is possible, after having launched the preceding request (*"selenophosphate synthetase"* AND (*"Homo Sapiens"* [organism])), to use in the **left panel** the filters *Molecule types* to keep only mRNA.

Question 19

Run the previous query again.
In the left panel, specify Molecule Type and click on mRNA to check this type of molecules.
Finally, link to the database *Genes*
How many entries do you get then?