# A Guide to Navigating GenBank and BLASTn Online

This is Part One of a two-part tutorial on using NCBI's databases. For the rest of the tutorial, please refer to the associated Jupyter Notebook *intro_to_NCBI.ipynb*.

NCBI provides multiple databases of genetic sequences, scienific articles and taxonomy data. The most interesting ones for your work at Genorobotics are the *Nucleotide* database for comparing your sample to all plant sequences on GenBank, and eventually the *Taxonomy* database for understanding the organization of the group of green plants: *Viridiplantae*.

NCBI's *Nucleotide* (also called Genbank) can be used to extract specific sequences or filtered databases. Each GenBank entry is characterized by a unique ID. As well as the species and the genetic sequence, the entry contains additional information like the authors, the location of the sample, etc.

**Note on IDs:** Historically, two IDs were used on GenBank to identify entries. The GenInfo Identifier (GI) and the accesion identifier (also sometimes denoted by "gb" in FASTA). Nowadays, the accesion identifier which contains both digits and letters is the preferred way to access entries. It is also uniformized across all genetic databases to ensure consistency, contrary to the GI. For more information on these two IDs: https://www.ncbi.nlm.nih.gov/genbank/sequenceids/

## Accessing a specific entry

Access the database through this link: https://www.ncbi.nlm.nih.gov/genbank/ and search for the sequence with the ID OQ303562.1

1- What's the scientific name of the organism? What about its common name?

**Note:** the scientific name is called the *bionomial nomenclature*, the first name is the genus (group of genetically related organisms with common ancestry) and the second is the specific species. For example: in Homo Sapiens, *Sapiens* is the species that qualifies modern humans and *Homo* is a larger group with other species like *Homo Erectus* and *Homo Abilis*.

2- Click on the organism name and hover over the lineage. What's the *family* of this species?

3- Go back to the entry's main page. Download the FASTA file with the sequence: on the top left → FASTA → Send to → File → Create file. Read the description string of the file, beginning with a ">". What information does it contain, which does it not?

**Note:** Here is a list of the most common keywords found in fasta file description strings for our genes of interest and their meaning:

| keyword | meaning |
|---|---|
| sp. | species without name, probably new |
| var. | Variant, sub-species |
| uncultured | mixed environmental sample (ex: soil) not sample from specific plant |
| unverified | GenBank hasn't verified the translation into protein |
| cf. | sequencer is not sure of species of sample |
| aff. | sample's species is not defined with certainty so the one mentioned is the closest morphologically. |

## Downloading Databases

In Genorobotics, the goal is to perform species identification **offline and on-site**, hence the need of downloading entire databases of genetic sequences to compare against. Here is an example. We will download the database for the gene MatK, one of the genes we use for species identification .

- Go back to Genbank's home page: https://www.ncbi.nlm.nih.gov/genbank/
- Go to Advanced search and set the following filters:
  - Gene Name: Matk
  - Sequence Length: 750:1500
- Download the entriesin a FASTA file: Send to → Complete Record → File → Format = Fasta → Create File. how big is the database you downloaded?
- Explore over other fields by which to filter the database and notice how filling these fields *changes the built search query* (what appears in the search bar). Understanding this search query structure will be essential to build code that automatically downloads databases as we will see in the Part 2 of the tutorial on the Jupyter notebook.

## Using BLASTn online

BLAST is a highly efficient alignment algorithm that can be used to compare two sequences of DNA or protein. It can be used online by providing a query sequence, which will get compared to the entire NCBI database, or  download it as an off-line tool that can be run on any database (cf. The installation tutorial on the Identification team's github).

For an overview of how BLAST works and why it is faster than previously explored algorithms like Needleman-Wunsch: https://en.wikipedia.org/wiki/BLAST_(biotechnology) (not very important)

BLASTn is the application of the use of this algorithm on the nucleotide database.

- Go to NCBI's BLAST page: https://blast.ncbi.nlm.nih.gov/Blast.cgi and pick the Nucleotide BLAST.
- With this tutorial comes a sequence file: my_sequence.fasta. Upload the file online or copy paste the DNA sequence. Run BLAST with the default option "Standard Databases"
- What are the best matches returned by BLAST, what is the percent identity and e-value?

**Note:** the percent identity is the percentage of identical bases. The e-value depends on different parameters used for alignment and thus constitutes a standardized way to evaluate the quality of an alignment. For e-values lower than $10^{-6}$, sequences can be considered identical.